A.A. Samarski

Introducción a los métodos numéricos

Editorial Mir Moscú

Introducción a los métodos numéricos

Introducción a los métodos numéricos

А. А. Самарский

Введение в численные методы

Издательство «Наука»

A. A. Samarski

Introducción a los métodos numéricos



Editorial Mir
Moscú

Traducido del ruso por el ingeniero K. P. Medkov

Impreso en la URSS

На испанском языке

Contenido

Capítulo IV
Métodos de diferencias de la resolución de los problemas de contorno para ecuaciones diferenciales ordinarias

Capítulo V
Problema de Cauchy para las ecuaciones diferenciales ordinarias

Capítulo VI
Métodos de diferencias para las ecuaciones *elípticas*

Capítulo VII
Métodos de diferencias para resolver la ecuación de conductibilidad térmica

Prólogo

Este libro representa una introducción a la teoría de los métodos numéricos en la que se emplea un mínimo de información de tales apartados de las matemáticas como son el análisis, el álgebra lineal y la teoría de ecuaciones diferenciales. El libro ha surgido como resultado de elaboración de las conferencias dictadas por el autor durante varios años para los estudiantes de la facultad de matemática de cálculo y cibernética de la Universidad de Moscú Lomonósov.

El contenido del libro es tradicional: interpolación y aproximación, integración numérica, resolución de ecuaciones no lineales, métodos directos e iterativos de resolución de los sistemas de ecuaciones algebraicas lineales, métodos de diferencias destinados a resolver el problema de Cauchy y problemas de contorno para las ecuaciones diferenciales ordinarias.

La aspiración del autor fue hacer la exposición comprensible de la primera lectura, prestando una atención especial a los conceptos principales de la teoría de los métodos numéricos e ilustrándolos con los ejemplos más simples.

Para la resolución numérica de varios problemas de la física y de la técnica descritos por las ecuaciones de la física matemática se emplea actualmente el método de diferencias finitas. Los conceptos principales de la teoría de los métodos de diferencias (aproximación, estabilidad, convergencia) se ilustran con ejemplos de esquemas de diferencias para las ecuaciones diferenciales ordinarias. Al aproximar las ecuaciones diferenciales, obtenemos ecuaciones en diferencias que representan sistemas de ecuaciones lineales de orden superior con matrices del tipo especial (tienen muchos elementos nulos), por ejemplo, tridiagonales. Un papel de

importancia lo desempeña la elección de los métodos efectivos (directos e iterativos) para resolver los sistemas mencionados. Con este motivo en el libro se exponen los fundamentos de la teoría general de métodos iterativos. Una gran atención se ha dedicado a la cuestión de estabilidad de los *cálculos en los ordenadores*. En el capítulo V viene una exposición sencilla de la teoría de estabilidad del problema de Cauchy para el sistema de ecuaciones en diferencias de primer orden. Aquí se han obtenido las condiciones coincidentes de estabilidad *necesarias y suficientes de los esquemas* de diferencias y, además, se ha investigado la estabilidad asintótica de los esquemas de diferencias.

En los dos últimos capítulos del libro (VI y VII) se analizan métodos de diferencias para resolver las ecuaciones elípticas y la ecuación de conductibilidad térmica. Estos capítulos son complementarios y permiten realizar el paso a la teoría de esquemas de diferencias para las ecuaciones en derivadas parciales.

Una exposición más detallada de los apartados separados de los métodos numéricos se da en los libros: «Teoría de esquemas de diferencias» por Samarski A. A., «Métodos de resolución de las ecuaciones reticulares» por Samarski A. A., Níkoláev E. S., y otros que se indican en la lista al final del libro.

El libro está destinado a los estudiantes de los primeros años que eligen como su especialidad la matemática aplicada y la física matemática; este libro puede resultar útil también para postgraduados y colaboradores científicos que estudien los métodos numéricos.

A. A. Samarski

Introducción

La aparición y el perfeccionamiento incesante de los ordenadores de alta velocidad han conducido a una transformación auténticamente revolucionaria de la ciencia en general y de las matemáticas, en particular. Ha cambiado la tecnología de las investigaciones científicas, han aumentado inmensamente las posibilidades de los estudios teóricos, del pronóstico de procesos complejos, de la proyección de las construcciones de ingeniería. Unicamente gracias a la aplicación de la simulación matemática y de nuevos métodos numéricos destinados para los ordenadores se hizo posible resolver grandes problemas científico-técnicos tales como el dominio de la energía nuclear y la asimilación del cosmos.

El primer gran problema, el dominio de la energía nuclear, requiere que se resuelva un conjunto de problemas complejos de la física y mecánica (manejo del trabajo de la caldera nuclear, la utilización de la energía proveniente de la fisión de los núcleos de uranio, la protección de la irradiación penetrante, el enfriamiento de las paredes de reactor, el estudio de los campos térmicos y de tensiones elásticas en las paredes, la resolución de varios otros problemas). Todos estos problemas han de ser resueltos antes de que empiece a trabajar una caldera, usando para este fin la descripción matemática (un modelo) y realizando cálculos numéricos en el ordenador. El segundo gran problema consistente en la asimilación del cosmos está relacionado con la creación de aparatos voladores y la resolución para estos últimos de diferentes problemas aerodinámicos y balísticos (por ejemplo, el cálculo del movimiento de un cohete y la dirección de su vuelo). En este dominio también hay un conjunto de problemas complejos de la mecánica, física

y técnica los cuales pueden ser resueltos sólo aplicando los **métodos** *numéricos*.

Indiquemos un problema más planteado ante la humanidad, esto es, la búsqueda de nuevas fuentes de energía. *Uno de los proyectos fundamentales para obtener energía* consiste en emplear la reacción de fusión termonuclear dirigida de los núcleos de deuterio y de tritio. Los recursos de combustible termonuclear en la Tierra son prácticamente inagotables, mientras que los productos de reacción no ensucian el ambiente. No obstante, la reacción termonuclear comienza sólo en condiciones extremadas: a una altísima temperatura (decenas y centenas de millones de grados) y enorme compresión (miles de veces) del deuterio y tritio; además, se requiere mantener la sustancia combustible en dicho estado durante un período de tiempo que sea suficiente para que se desarrolle la *reacción* de combustión (del síntesis). La creación de las condiciones mencionadas es un problema científico-técnico que por ahora no está resuelto. *Existen varios proyectos destinados a calentar*, comprimir y mantener el combustible termonuclear (plasma). Al realizarlos surge una serie de cuestiones que *deben ser resueltas antes de proceder* a la proyección de las instalaciones correspondientes, incluso experimentales. Es menester estudiar ante todo el comportamiento del plasma a altas temperaturas y densidades, en campos magnéticos y, además, aclarar las condiciones bajo las cuales resulta posible la propia reacción de la síntesis termonuclear.

Las investigaciones de tal índole se efectúan a base de la descripción matemática (modelo matemático) de los procesos físicos y la resolución ulterior de problemas matemáticos correspondientes en el ordenador con ayuda de algoritmos de cálculo (computacionales).

Hoy día podemos decir que ha surgido un método nuevo para la investigación teórica de los procesos complejos que admiten la descripción matemática: se trata de un experi**mento de cálculo**, *es decir, la investigación de los problemas* científicos naturales por medio de la matemática de cálculo. Expliquemos la esencia de este método de investi*gación con un ejemplo de resolución* de un problema físico. Supongamos que se pide estudiar cierto proceso físico. A la investigación matemática le precede la elección de una

aproximación física, es decir, se debe determinar qué factores han de tomarse en consideración y cuáles pueden ser menospreciados. Resuelta la cuestión citada, se realiza la investigación del problema mediante un experimento de cálculo, en el que pueden distinguirse las siguientes etapas principales.

En la primera etapa se elige un modelo matemático, es decir, la descripción aproximada del proceso en forma de ecuaciones algebraicas, diferenciales o integrales. Estas ecuaciones expresan corrientemente las leyes de conservación de las magnitudes físicas principales (la energía, la cantidad de movimiento, la masa, etc.). El modelo matemático obtenido ha de ser investigado recurriendo a la teoría de ecuaciones diferenciales. Se debe establecer si el problema está planteado correctamente, si los datos de partida son suficientes y ellos no contradicen los unos a los otros, si existe la solución del problema planteado y si es única. En esta etapa se emplean los métodos de la matemática clásica. Hemos de señalar que muchos problemas físicos conducen a ciertos modelos matemáticos cuya elaboración teórica acaba de iniciarse. En la práctica nos vemos obligados a resolver problemas de la física matemática, para los cuales no existen teoremas de existencia y unicidad.

La segunda etapa del experimento de cálculo consiste en la construcción de un método numérico aproximado que se usa para resolver el problema, es decir, en la elección del algoritmo de cálculo. Por algoritmo de cálculo se entiende una sucesión de operaciones aritméticas y lógicas que ayudan a encontrar la solución del problema matemático formulado en la primera etapa. Más abajo se discutirán detalladamente las exigencias que se presentan a un algoritmo de cálculo destinado para el empleo en los ordenadores modernos. El presente libro está dedicado, en esencia, al estudio de los algoritmos de cálculo elementales.

En la tercera etapa se lleva a cabo la programación del algoritmo de cálculo para el ordenador y en la cuarta etapa, los cálculos en el ordenador. No nos detendremos en las cuestiones ligadas con la programación, organización y realización de los cálculos en el ordenador, puesto que todas estas cuestiones salen de los márgenes del libro. Notemos sólo que todas las operaciones referentes a la programación

deben estar en una relación estrecha con la elaboración de los algoritmos numéricos concretos.

En fin, a título de la quinta etapa del experimento de cálculo puede indicarse el análisis de los resultados numéricos obtenidos y la precisión posterior del modelo matemático. Puede suceder que el modelo es demasiado aproximado (el resultado de los cálculos no concuerda con el experimento físico) o bien el modelo es muy complejo y la solución puede obtenerse con una exactitud suficiente empleando modelos más simples. En este caso el trabajo se inicia desde la primera etapa, es decir, se precisa el modelo matemático y se repasan otra vez todas las etapas.

Hemos de notar que un experimento de cálculo no es, como regla, una operación sencilla de cálculo por fórmulas estándar, sino, ante todo, los cálculos de toda una serie de variantes para diferentes modelos matemáticos.

Ahora, fijemos nuestra atención en ciertas características y exigencias generales concernientes a los algoritmos de cálculo. La elaboración e investigación de los algoritmos de cálculo y la aplicación de éstos a la resolución de los problemas concretos constituyen el contenido de un gran apartado de la matemática moderna: matemática de cálculo.

La matemática de cálculo se determina en el amplio sentido de este término como un apartado de las matemáticas que incluye un conjunto de cuestiones relacionadas con el empleo de los ordenadores; en el sentido estrecho dicho apartado de las matemáticas se entiende como teoría de los métodos numéricos y de los algoritmos para resolver los problemas matemáticos planteados. En lo sucesivo la matemática de cálculo se considerará sólo en el sentido estrecho.

Hay un rasgo común para todos los métodos numéricos que consiste en reducir todo problema matemático a uno que sea de dimensión finita. Esto se consigue con mayor frecuencia discretizando el problema de partida, es decir, pasando de las funciones de un argumento continuo a las de argumento discreto. Discretizado el problema de partida, se debe construir un algoritmo de cálculo, es decir, indicar la sucesión de operaciones aritméticas y lógicas que se ejecutan en el ordenador y que proporcionan, tras un número finito de operaciones, la solución del problema discreto.

La solución obtenida del problema discreto se considera como solución aproximada del problema matemático de partida.

Al resolver problemas en el ordenador obtenemos siempre no la solución exacta del problema de partida, sino cierta solución aproximada. ¿A qué se debe el error que surge? Pueden ser indicadas tres razones principales a consecuencia de las cuales surgen errores en la resolución numérica del problema matemático de partida. Ante todo, los datos de entrada del problema de partida (condiciones iniciales y de frontera, coeficientes y segundos miembros de las ecuaciones) se dan siempre con cierta inexactitud. Un error del método numérico condicionado por la prefijación inexacta de los datos de entrada suele denominarse *error inevitable*. Luego, al sustituir el problema de partida por otro problema discreto aparece un error que se llama *error de discretización* o, de otra forma, *error del método*. Por ejemplo, sustituyendo la derivada $u'(x)$ por una razón de diferencias $(u(x + \Delta x) - u(x))/\Delta x$, cometemos un error de discretización que para $\Delta x \to 0$ tiene el orden Δx. Finalmente, el orden finito de los números que se suministran al ordenador lleva a *errores de redondeo* que pueden acumularse en el transcurso de los cálculos. Es natural exigir que los errores en la prefijación de la información inicial y el error que surge como resultado de discretización sean concordados con el error de la solución del problema discreto en el ordenador.

De lo dicho proviene que la exigencia principal que se levanta ante el algoritmo de cálculo es exactitud. Dicha exigencia quiere decir que el algoritmo de cálculo debe asegurar la solución del problema de partida con la *exactitup* prefijada $\varepsilon > 0$, realizadas un número finito $Q(\varepsilon)$ de operaciones. El algoritmo ha de ser realizable, es decir, debe proporcionar la solución del problema en tiempo de máquina admisible. Para la mayoría de los algoritmos el tiempo que se necesita para resolver el problema (volumen de los cálculos) $Q(\varepsilon)$ crece al aumentar la exactitud, es decir, cuando disminuye ε. Por supuesto, se puede prefijar ε tan pequeño que el tiempo de resolución del problema se hará inadmisiblemente grande. Resulta importante conocer que el algoritmo da en principio una posibilidad de obtener la solución del problema con cualquier exactitud. Sin embargo, en la

práctica la magnitud de ε se elige tomando en consideración una posibilidad de realizar el algoritmo en el ordenador dado. Para cualquier problema, algoritmo y *ordenador* existe un valor individual de ε.

Es natural de aspirar a que el número de operaciones (y, de este modo, el tiempo de máquina para la resolución del problema) $Q(\varepsilon)$ sea mínimo para el problema dado. Para cualquier problema se pueden ofrecer varios algoritmos que proporcionen (para $\varepsilon \to 0$) una exactitud $\varepsilon > 0$ *igual* en orden, pero con diferente número de operaciones $Q(\varepsilon)$. Entre estos algoritmos (de los cuales suele decirse que ellos son equivalentes según el orden de exactitud) se *debe elegir uno que proporcione la solución* con un gasto mínimo de tiempo de máquina (número de operaciones $Q(\varepsilon)$). Tales algoritmos se denominarán *económicos*.

He aquí una exigencia más que ha de ser *satisfecha por* el algoritmo de cálculo, es decir, el requisito de que no haya parada de emergencia (de indisponibilidad) del ordenador en el proceso de los cálculos.

Es necesario tener en cuenta que todo ordenador opera con números que tienen una cantidad finita de cifras significativas y que pertenecen (en módulo) no a todo el eje numérico, sino a cierto intervalo (M_0, M_∞), $M_0 > 0$, $M_\infty < \infty$, donde M_0 es un cero de máquina y M_∞, un infinito de máquina. Si la condición $|M| < M_\infty$ no se *cumple en el proceso de los cálculos*, ocurre una parada de emergencia del ordenador («parem»), a consecuencia de que queda rellenada la red de órdenes y los cálculos se dan por terminado. La posibilidad de una *parem depende tanto* del algoritmo como del problema de partida.

Si la solución del problema de partida se expresa en términos de números muy grandes (muy pequeños) $|M| > M_\infty$ ($|M| < M_0$), entonces, como regla, variando la escala, el problema puede ser reducido a una forma que contiene sólo las magnitudes pertenecientes (en módulo) al intervalo prefijado (M_0, M_∞). La posibilidad de la parem se elimina frecuentemente cambiando el orden de operaciones. Expliquémoslo con un ejemplo sencillo.

EJEMPLO. *Sea* $M_\infty = 10^p$, $M_0 = 10^{-p}$, $p = 2^n$, *n es* un número entero. Se pide calcular el producto de los números $10^{p/2}$, $10^{p/4}$, $10^{-p/2}$, $10^{3p/4}$, $10^{-3p/4}$.

1er MÉTODO. Fijemos los números en el orden decreciente:

$$q_1 = 10^{3p/4}, \qquad q_2 = 10^{p/2}, \qquad q_3 = 10^{p/4},$$
$$q_4 = 10^{-p/2} \qquad q_5 = 10^{-3p/4}$$

y formemos los productos $S_{k+1} = S_k q_{k+1}$, $S_1 = q_1$. En este caso, ya en el primer paso tendrá lugar una parem, puesto que $S_2 = q_1 q_2 = 10^{5p/4} > M_\infty$.

2do MÉTODO. Fijemos los números en el orden creciente:

$$q_1 = 10^{-3p/4}, \qquad q_2 = 10^{-p/2}, \qquad q_3 = 10^{p/4},$$
$$q_4 = 10^{p/2}, \qquad q_5 = 10^{3p/4}.$$

En este caso obtendremos en el primer paso

$$S_2 = q_1 q_2 = 10^{-5p/4} < M_0,$$

es decir, S_2 es un cero de máquina; todos los productos sucesivos S_3, S_4, S_5 son también nulos; de este modo aquí ocurre una pérdida total de exactitud.

3er MÉTODO. Mezclemos estos números suponiendo $q_1 = 10^{-3p/4}$, $q_2 = 10^{p/2}$, $q_3 = 10^{3p/4}$, $q_4 = 10^{-p/2}$, $q_5 = 10^{p/4}$. Entonces hallaremos sucesivamente:

$$S_2 = q_1 q_2 = 10^{-p/4}, \qquad S_3 = S_2 q_3 = 10^{p/2},$$
$$S_4 = S_3 q_4 = 10^0, \qquad S_5 = S_4 q_5 = 10^{p/4},$$

es decir, en el proceso de los cálculos no aparecen números superiores a $10^{p/2}$ e inferiores a $10^{-p/4}$. Tal algoritmo está privado de parem. En el cap. III nos encontraremos con un método iterativo de resolución de sistemas de ecuaciones algebraicas lineales que puede realizarse con una parem y sin ésta, según sea la forma de numeración de los parámetros que determina la sucesión de los cálculos.

En cada etapa de los cálculos surgen errores de redondeo. Estos errores de redondeo pueden crecer o ir disminuyendo, en dependencia del algoritmo.

Si, en el transcurso de los cálculos, la magnitud de los errores de redondeo crece indefinidamente, el algoritmo se llamará *inestable* (desde el punto de vista de cálculos). En cambio, si los errores de redondeo no se acumulan, el algoritmo será estable.

EJEMPLOS. 1. Supongamos que se pide hallar y_i $(0 < i \leqslant i_0)$ según la fórmula $y_{i+1} = y_i + d$ $(i \geqslant 0)$ para y_0, d prefijados. Supongamos, además, que al calcular y_i se ha introducido un error (por ejemplo, un error de redondeo) cuya magnitud es δ_i, es decir, en lugar del valor exacto de y_i tenemos un valor aproximado $\tilde{y}_i = y_i + \delta_i$. Entonces, en vez del valor exacto de y_{i+1} obtendremos el valor aproximado $\tilde{y}_{i+1} = (\tilde{y}_i + \delta_i) + d = y_{i+1} + \delta_i$. De este modo, un error cometido en cualquier paso intermedio no aumenta en el proceso de los cálculos. El algoritmo es estable.

2. Examinemos la ecuación $y_{i+1} = qy_i$ $(i \geqslant 0$, y_0 y q están prefijados). Supongamos, al igual que en el ejemplo 1, se ha obtenido, en lugar de y_i, el valor $\tilde{y}_i = y_i + \delta_i$. Entonces, en lugar de y_{i+1} obtendremos un valor aproximado

$$\tilde{y}_{i+1} = q\,(y_i + \delta_i) = y_{i+1} + q\delta_i.$$

De aquí se ve que el error $\delta_{i+1} = \tilde{y}_{i+1} - y_{i+1}$, que surge al calcular y_{i+1}, está ligado con el error δ_i mediante una ecuación

$$\delta_{i+1} = q\delta_i, \qquad i = 0,\ 1,\ 2,\ \ldots$$

Por consiguiente, si $|\,q\,| > 1$, el valor absoluto del error crecerá en el proceso de los cálculos (el algoritmo es inestable). Si, en cambio, $|\,q\,| \leqslant 1$, entonces el error no aumenta, es decir, el algoritmo es estable. La inestabilidad se liga corrientemente con la propiedad de crecimiento exponencial del error de redondeo. Si el error de redondeo crece según la ley potencial al pasar de una operación a la otra («de paso a paso»), el algoritmo se considera *convencionalmente estable* (estable con ciertas restricciones que se imponen sobre el volumen de cálculos y la exactitud requerida). El proceso de los cálculos puede interpretarse así: al pasar de un paso a otro tiene lugar una alteración (a cuenta de los errores de redondeo) de las últimas cifras significativas («una onda del error de redondeo» se mueve de derecha a izquierda, partiendo de las últimas cifras significativas). Nuestra tarea consiste en conservar justas unas cuantas primeras cifras significativas (4—5 signos) y por esta razón los cálculos deben darse por terminado antes de que «la

onda del error de redondeo» alcance dichas cifras. Si el error de redondeo ε_0 crece de un paso al otro según la ley exponencial, esto conduce, como regla, a una parem en cierta etapa intermedia de los cálculos, si (lo mismo que en el ejemplo 2) $|q|^i\varepsilon_0 \geqslant M_\infty$.

Si $M_\infty = 10^p$, $\varepsilon_0 = 10^{-k_0}$, la parem llega cuando $i_0 > (p + k_0)/\lg|q|$. Otra cosa ocurre cuando el error de redondeo crece según la ley potencial. Sea $|\delta y_i| \approx i^n\varepsilon_0$ $(n \geqslant 1)$; entonces, la parem tiene lugar para $i_0^n\varepsilon_0 \geqslant M_\infty$, es decir, para $i_0 \geqslant \left(\frac{1}{\varepsilon_0}M_\infty\right)^{1/n} = 10^{(p+k_0)/n}$.

De aquí se ve que para $n = 1$ la parem no tendrá lugar en virtud de una restricción evidente $i < M_\infty = 10^p$. La desigualdad $|\delta y_i| \leqslant \varepsilon$, donde $\varepsilon = 10^{-k}$ es la exactitud prefijada, se verifica para $i \leqslant \left(\frac{\varepsilon}{\varepsilon_0}\right)^{1/n} = 10^{(k_0-k)/n} = i_0$. Si están prefijados ε y ε_0, esta desigualdad significa una restricción para el número de ecuaciones $i \leqslant i_0$. Por ejemplo, para $k_0 = 12$, $k = 6$, tenemos $i \leqslant 10^{6/n}$, de suerte que $i \leqslant 10^3$ para $n = 2$. Está claro que puede elegirse tal n grande que el número admisible de ecuaciones i_0 sea muy pequeño. Sin embargo, en la práctica se encuentran corrientemente casos de n pequeño (por ejemplo, para el método de factorización (§ 3 cap. I) $n = 2$, es decir, el error se acumula según la ley cuadrática a medida que crece el número de ecuaciones).

Al resolver un problema (cualquiera que sea) es necesario conocer ciertos datos de entrada (de partida): datos iniciales, valores de frontera de la función buscada, coeficientes y el segundo miembro de la ecuación, etc.

Para todo problema se buscan respuestas a las preguntas de un mismo género: si existe la solución del problema, si será única y cómo depende la solución de los datos de entrada. Son posibles dos casos:

El problema está correctamente planteado (es correcto); esto quiere decir que: 1) el problema es resoluble para cualesquiera datos de entrada admisibles; 2) se tiene una única solución; 3) la solución del problema depende continuamente de los datos de entrada (a una variación pequeña de los datos de entrada le corresponde una variación pequeña de la solución), en otras palabras, el problema es estable.

El problema no está correctamente planteado (no es correcto), si la solución de éste es inestable respecto a los datos de entrada (a una variación *pequeña de los datos de* entrada le puede corresponder una variación grande de la solución).

Como ejemplo de un problema correcto puede servir el problema de integración y como ejemplo de un problema no correcto, el problema de diferenciación.

EJEMPLOS. 1. PROBLEMA DE INTEGRACIÓN. Sea dada una función $f(x)$; hállese la integral

$$J=\int_0^1 f(x)\,dx.$$

Sustituyamos f por $\tilde{f}$ y veamos $\tilde{J}=\int_0^1 \tilde{f}(x)\,dx$ y la diferencia $\delta J=\tilde{J}-J=\int_0^1 \delta f\,dx\ (\delta f=\tilde{f}(x)-f(x))$. De aquí se ve que $|\delta J|\leqslant \max\limits_{0\leqslant x\leqslant 1}|\delta f(x)|$, $|\delta J|\leqslant\varepsilon$, si $|\delta f|\leqslant\varepsilon$, es decir J depende *continuamente de f. Con el fin de calcular* la integral J hagamos uso de la fórmula de cuadratura:

$$J_N=\sum_{k=1}^{N} c_k f(x_k),\quad c_k>0,\ \sum_{k=1}^{N} c_k=1.$$

Al repetir los razonamientos aducidos más arriba, llegamos a que

$$\delta J_N=\tilde{J}_N-J_N=\sum_{k=1}^{N} c_k(\tilde{f}_k-f_k)=\sum_{k=1}^{N} c_k\delta f_k,$$

$$|\delta J_N|\leqslant\sum_{k=1}^{N} c_k\max_{1\leqslant k\leqslant N}|\delta f_k|=\max_{1\leqslant k\leqslant N}|\delta f_k|.$$

De este modo, el problema de cálculo de una integral por la fórmula de cuadratura es correcto.

2. PROBLEMA DE DIFERENCIACIÓN. El problema de diferenciación de una función $u(x)$ definida aproximadamente no es correcto.

En efecto, sea $\tilde{u}(x) = u(x) + \frac{1}{N} \operatorname{sen} N^2 x$, donde N es suficientemente grande. Entonces, en la métrica C (en cierto segmento $0 \leqslant x \leqslant \delta$ ($\delta > \pi/N^2$)) tenemos $\| \delta u \|_C = \| \tilde{u} - u \|_C = 1/N \leqslant \varepsilon$ para $N \geqslant 1/\varepsilon$. Para el error de las derivadas $\delta u' = \tilde{u}' - u' = N \cos N^2 x$ tenemos $\| \delta u' \|_C = N \geqslant 1/\varepsilon$. De este modo, a una variación pequeña $O(\varepsilon)$ en C de la función $u(x)$ le corresponde la variación grande $O(1/\varepsilon)$ en C de su derivada.

Por eso la diferenciación numérica tampoco es correcta. Para encontrar un valor aproximado de una derivada según la fórmula de la derivada de diferencias con cierta exactitud $\varepsilon > 0$ a condición de que la función viene definida con un error δ_i ($| \delta_i | \leqslant \delta_0$), hace falta que se cumplan las condiciones de concordancia entre ε, δ_0 y el paso h de la red, por ejemplo, del tipo $\varepsilon \geqslant k \sqrt{\delta_0}$ ($k = \text{const} > 0$ no depende de h, δ_0), con la particularidad de que el paso de la red está acotado tanto inferiormente como superiormente. De este modo, la exactitud alcanzable de la diferenciación numérica está limitada por la exactitud con la que viene definida la propia función.

En este libro se estudian sólo problemas correctos y métodos numéricos correctos destinados para aplicarlos en el ordenador.

Los métodos numéricos dan la solución aproximada del problema. Esto significa que en lugar de la solución exacta u (de una función o de una funcional) de algún problema encontramos una solución y de otro problema, próxima en cierto sentido (según la norma, por ejemplo) a la buscada. De acuerdo con lo dicho, la idea principal de todos los métodos consiste en discretizar o aproximar (sustitución, aproximación) el problema de partida a algún otro que sea más cómodo para resolverlo en un ordenador, con la particularidad de que la resolución del problema que aproxima depende de ciertos parámetros que, siendo manejados de una manera adecuada, permiten determinar la solución con una exactitud requerida. Por ejemplo, en el problema de integración numérica los nodos y los pesos de la fórmula de cuadratura representan precisamente los parámetros de este tipo. Luego, la solución de un problema discreto es elemento de

un espacio de dimensión finita. Expliquemos esto más detalladamente.

Veamos, por ejemplo, la discretización de un espacio $H = \{f(x)\}$ de funciones $f(x)$ de argumento continuo $x \in [a, b]$. *Introduzcamos en un segmento* $a \leqslant x \leqslant b$ un conjunto finito de puntos $\omega = \{x_i, i = 0, 1, \ldots, N, x_0 = a, x_N = b, x_i < x_{i+1}\}$ el cual se llamará *red*. Los *puntos* x_i se denominarán *nodos* de la red ω. Si la distancia $h_i = x_i - x_{i-1}$ entre los nodos vecinos es constante (no depende de i), $h_i = h$ para todos los $i = 1, 2, \ldots, N$, entonces la red se denomina *uniforme* (de *paso* h); en caso contrario se denomina *no uniforme*. *En lugar de la función* $f(x)$ definida para cualesquiera $x \in [a, b]$ consideraremos una *función reticular* $y_i = f(x_i)$ de argumento i $(i = 0, 1, \ldots, N)$, que es un número entero, o bien de nodo x_i de la red ω, y sustituiremos $H = \{f(x), x \in [a, b]\}$ por un espacio de *dimensión finita* (de dimensión $N + 1$) $H_{N+1} = \{y_i, 0 \leqslant i \leqslant N\}$ de funciones reticulares. Es evidente, que la función reticular $y_i = f(x_i)$ puede *considerarse como* un vector $y = (y_0, y_1, \ldots, y_N)$.

Podemos discretizar también el espacio de funciones $f(x)$ de varias variables, si $x = (x_1, x_2, \ldots, x_p)$ es un punto del espacio euclídeo p-dimensional $(p > 1)$. Por ejemplo, *en un plano* (x_1, x_2) *se puede introducir* una red $\omega = \{x_i = (i_1 h_1, i_2 h_2), i_1, i_2 = 0, \pm 1, \pm 2, \ldots\}$ como conjunto de puntos (nodos) de intersección de las rectas perpendiculares $x_1^{(i_1)} = i_1 h_1$, $x_2^{(i_2)} = i_2 h_2$, $h_1 > 0$, $h_2 > 0$, $i_1, i_2 = 0, \pm 1, \pm 2, \ldots$, donde h_1 y h_2 son los pasos de la red según las direcciones de x_1 y x_2, respectivamente. La red ω es, evidentemente, uniforme según cada una de las *variables por separado*. En lugar de la función $f(x) = f(x_1, x_2)$ analizaremos una función reticular

$$y_{i_1 i_2} = f(i_1 h_1, i_2 h_2).$$

Si la red ω *contiene sólo los* nodos pertenecientes al rectángulo $(0 \leqslant x_1 \leqslant l_1, 0 \leqslant x_2 \leqslant l_2)$ de modo que $h_1 = l_1/N_1$, $h_2 = l_2/N_2$, entonces la red cuenta *con un número* finito $N = (N_1 + 1)(N_2 + 1)$ de nodos, mientras que el espacio H_N de funciones reticulares $y_i = y_{i_1, i_2}$ es de dimensión finita.

En todos los casos consideramos sólo un espacio de funciones reticulares cuya dimensión es finita. Al sustituir el espacio $H = \{f(x)\}$ de funciones de argumento continuo por el espacio H_N de funciones reticulares y el problema de partida, por su aproximación discreta, debemos estar seguros de que nos aproximamos mejor a la solución del problema de partida aumentando el número de nodos. La estimación de la calidad de aproximación y la elección del método de aproximación constituyen la tarea principal de la teoría de los métodos numéricos.

El contenido fundamental del libro está relacionado de una u otra manera con la aplicación de los métodos de diferencias para solucionar ecuaciones diferenciales. Destaquemos dos momentos de importancia:

— obtención de la aproximación discreta (de diferencias) de las ecuaciones diferenciales e investigación de las ecuaciones en diferencias que aparecen en este caso;

— resolución de las ecuaciones en diferencias.

Al obtener una aproximación discreta (esquema de diferencias) un papel importante lo desempeña el requisito general referente a que el esquema de diferencias aproxime al máximo (simule) las propiedades principales de la ecuación diferencial inicial. Tales esquemas de diferencias pueden obtenerse, por ejemplo, con ayuda de los principios variacionales y las relaciones integrales (véase el cap. IV). La estimación de la exactitud de un esquema de diferencias se reduce al estudio del error de aproximación y de la estabilidad del esquema. El estudio de la estabilidad es una cuestión central de la teoría de los métodos numéricos y a ella se le presta gran atención en el presente libro. Los algoritmos de los problemas complejos se pueden representar como una sucesión (cadena) de algoritmos simples (módulos). Por eso muchos problemas de principio de la teoría de los métodos numéricos pueden aclararse con algoritmos simples.

En el primer capítulo se examinan las ecuaciones en diferencias unidimensionales (que dependen de un argumento entero). Nos limitamos al estudio de las ecuaciones en diferencias de primero y segundo órdenes. Las ecuaciones en diferencias de segundo orden representan un sistema de ecuaciones algebraicas lineales con matriz tridiagonal. Con

el objeto de resolver los problemas de contorno para estas ecuaciones se emplea el así llamado método de factorización. En el primer capítulo se dan, a título de un material de información, algunos conocimientos sobre los operadores lineales en un espacio de dimensión finita. Posteriormente se investigan las propiedades de los operadores de diferencias en su calidad de operadores lineales en un espacio de dimensión finita provisto de un producto escalar. En este caso se emplea un aparato matemático más sencillo, es decir, las fórmulas para la diferenciación de diferencias de un producto y para la adición por partes.

En el segundo capítulo se expone el material tradicional del análisis numérico: la interpolación, la aproximación media cuadrática y la integración numérica.

Al aproximar las ecuaciones diferenciales en una red se obtienen ecuaciones en diferencias que representan un sistema de ecuaciones algebraicas lineales de orden superior (igual al número de nodos de la red) con una matriz especial (enrarecida, es decir, una matriz que tiene muchos elementos nulos). Un ejemplo más simple de tal matriz (una matriz tridiagonal) fue indicado anteriormente.

En el tercer capítulo se exponen los métodos numéricos de resolución de las ecuaciones algebraicas lineales

$$\sum_{j=1}^{N} a_{ij}u^j = f^i \qquad i = 1, 2, \ldots, N, \tag{1}$$

las cuales pueden ser escritas en forma matricial

$$Au = f, \tag{2}$$

donde $A = (a_{ij})$ es una matriz cuadrada de dimensión $N \times N$, $u = (u^1, u^2, \ldots, u^N)$ es el vector buscado y $f = (f^1, f^2, \ldots, f^N)$, el vector prefijado (el segundo miembro).

Para resolver los sistemas de ecuaciones se usan métodos directos e iterativos.

En el § 2 del cap. III se analizan el método de eliminación de Gauss y el de raíz cuadrada que representan métodos directos los cuales requieren $O(N^3)$ operaciones aritméticas para resolver el sistema.

Al estudiar los métodos iterativos, resulta cómodo interpretar el sistema de ecuaciones algebraicas lineales (2)

como ecuación operacional de primera especie con un operador que actúa en el espacio N-dimensional H_N (A: $H_N \to H_N$), u, $f \in H_N$. Para subrayar la equivalencia existente entre las formas de escritura matricial y operacional, la matriz y el operador correspondiente se designarán con una misma letra A.

En la teoría de los métodos iterativos (de un paso o de dos capas) es de mucha importancia la forma canónica del esquema iterativo

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\qquad k=0,\ 1,\ \ldots \text{ para todo } y_0\in H_N, \tag{3}$$

donde A, B: $H_N \to H_N$, $\{\tau_k\}$ son parámetros iterativos.

En cualquier caso se supone que el operador A es autoconjugado y definido positivo ($A = A^* > 0$). Está demostrado el *teorema general sobre la convergencia del método* estacionario con $\tau_k = \tau = \text{const}$. Como condición suficiente de la convergencia interviene una desigualdad

$$(By,\ y) > \frac{\tau}{2}(Ay,\ y) \text{ para todo } y \in H, \tag{4}$$

donde $B \neq B^*$ es, *en el caso general, un operador no autoconjugado*. De aquí se desprende la convergencia del método de iteración simple, del método de Seidel y del de relajación superior.

Si se conocen unas constantes $\gamma_1 > 0$, $\gamma_2 > \gamma_1$, tales que

$$\gamma_1 (Bx,\ x) \leqslant (Ax,\ x) \leqslant \gamma_2 (Bx,\ x) \text{ para cualquier } x \in H_N, \tag{5}$$

donde $B = B^* > 0$, entonces podemos encontrar una totalidad optimal de parámetros de Chébishev $\{\tau_k^*\}$, con los cuales el proceso de cálculo es estable y se realiza sin parem.

Se examina el método universal alternado triangular con la totalidad $\{\tau_k^\alpha\}$ y un operador

$$B = (D + \omega A_1)\, D^{-1} (D + \omega A_2), \tag{6}$$

donde $D = D^* > 0$, $A_1^* = A_2$, $A_1 + A_2 = A$, las matrices A_1 y A_2 son triangulares. Hemos obtenido la fórmula para

el parámetro ω. El algoritmo para este método es muy simple. En todo caso se dan a conocer las fórmulas para el número de iteraciones con las cuales se alcanza la exactitud requerida. Los diferentes métodos fueron comparados a base de un problema modelo para la ecuación en diferencias de segundo orden $y_{i-1} - 2y_i + y_{i+1} = -h^2 f_i$, $i = 1, 2, \ldots, N - 1$, $y_0 = y_N = 0$, $h = 1/N$, que corresponde al problema de contorno $u''(x) = -f(x)$ $(0 < x < 1)$, $u(0) = u(1) = 0$. Esta ecuación es un análogo unidimensional de la ecuación de Laplace. Por cuanto el número de iteraciones no depende prácticamente del número de mediciones, entonces en el proceso de comparación podemos limitarnos a este problema unidimensional. El método alternativo triangular exige $O\left(\frac{1}{\sqrt{h}} \ln \frac{1}{\varepsilon}\right)$ iteraciones, donde $\varepsilon > 0$ es la exactitud prefijada.

Ha de ser notado que en el cap. III, en forma lo suficientemente completa, está expuesta de hecho, con ayuda de los medios matemáticos más sencillos, la teoría general de los métodos iterativos para resolver la ecuación $Au = f$ $(A = A^* > 0)$.

Los conceptos fundamentales de la teoría de esquemas de diferencias —error de aproximación, estabilidad, convergencia y exactitud— se exponen a base de ejemplos de los problemas de contorno y del problema de Cauchy para ecuaciones diferenciales ordinarias (cap. IV y cap. V). En el cap. IV se analizan esquemas de diferencias tripuntuales para una ecuación diferencial ordinaria de segundo orden

$$\frac{d}{dx}\left(k(x)\frac{du}{dx}\right) - q(x)u = -f(x), \qquad 0 < x < 1,$$

$$u(0) = u_1, \quad u(1) = u_2, \quad k(x) > 0, \quad q(x) \geqslant 0. \qquad (7)$$

Se han investigado las cuestiones referentes a la velocidad de convergencia (orden de exactitud) de los esquemas homogéneos de diferencias sobre las redes no uniformes y para el caso de coeficientes discontinuos. Esto ha exigido que se obtengan estimaciones apriorísticas bastante finas que expresen la estabilidad del esquema de diferencias respecto al segundo miembro.

Para obtener los esquemas de diferencias pueden utilizarse algunos métodos más diferentes: integral de interpolación, de aproximación de la funcional cuadrática, los de Ritz y Galerkin (§ 5, cap. IV).

Con el fin de resolver el problema de Cauchy para la ecuación de primer orden

$$\frac{du}{dt}=f(t, u), \qquad t>0, \qquad u(0)=u_0 \qquad (8)$$

se emplean los métodos de Runge—Kutta y de Adams expuestos en el cap. V. Estos métodos son también aplicables para un sistema de ecuaciones en que f y u son vectores.

Una atención especial en el cap. V se presta al problema de Cauchy para el sistema de ecuaciones lineales

$$\frac{du}{dt}+Au=f(t), \qquad t>0, \qquad u(0)=u_0, \qquad (9)$$

donde $A=(a_{ij})$ es una matriz cuadrada $N\times N$, $u(t)=(u^1, u^2, \ldots, u^N)$, $f(t)=(f^1, f^2, \ldots, f^N)$ es una función vectorial de N-ésima dimensión.

Tal problema surge, en particular, si en la ecuación de conductibilidad térmica

$$\frac{\partial u}{\partial t}=\Delta u+f(x, t), \qquad \Delta u=\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}, \; x=(x_1, x_2) \qquad (10)$$

sustituimos el operador de Laplace Δu por el operador de diferencias correspondiente. Entonces, (9) puede interpretarse como un método de las rectas para la ecuación de conductibilidad térmica (10). Empleando para resolver este problema algún esquema de un paso, llegamos a ún esquema operacional de diferencias de dos capas de forma general, el cual se escribe en la forma canónica

$$B\,\frac{y_{k+1}-y_k}{\tau}+Ay_k=\varphi_k, \qquad k=0, 1, \ldots,$$
$$\text{para todo } y_0\in H_N, \qquad (11)$$

donde A, B: $H_N\to H_N$ son los operadores lineales, τ es el paso de la red según t.

Se ha demostrado que la condición necesaria y suficiente de estabilidad del esquema tiene por expresión

$$B \geqslant \frac{\tau}{2} A, \text{ o bien } (Bx, x) \geqslant \frac{\tau}{2}(Ax, x)$$

$$\text{para todo } x \in H_N. \qquad (12)$$

Este es el teorema fundamental de la teoría general de estabilidad de esquemas operacionales de diferencias (véase «Teoría de los esquemas de diferencias» por Samarski A.A.) aplicable en la investigación de la estabilidad de esquemas de diferencias para las ecuaciones con derivadas parciales de la física matemática (véase el cap. VII). En realidad, en el § 4 están expuestos los fundamentos de la teoría general de estabilidad de los esquemas de diferencias, incluida la estabilidad asintótica.

Los conocimientos dados a conocer en los capítulos III, IV y V permiten pasar sin dificultad alguna al estudio de la teoría de los métodos de diferencias para resolver ecuaciones en derivadas parciales. En el cap. VI este estudio se ha realizado para esquemas de diferencias que aproximan la ecuación de Poisson y las ecuaciones elípticas en un rectángulo con condiciones de contorno de primera especie. Aquí están analizadas tanto las cuestiones de convergencia como los métodos de resolución de las ecuaciones en diferencias.

La teoría general de estabilidad de los esquemas de diferencias de dos capas (cap. V) simplifica la exposición de los métodos de diferencias para la ecuación de conductibilidad térmica con coeficientes constantes y variables realizada en el cap. VII. En el mismo capítulo se analizan también los esquemas económicos (de direcciones variables, de fisión, etc.) para los problemas multidimensionales, como también el principio general de la aproximación sumaria el que permite efectuar la partición de los problemas complejos en una sucesión de problemas más sencillos y debido a ello simplificar considerablemente la resolución de los problemas multidimensionales de la física matemática.

Se debe observar que el contenido principal de este libro se expone desde un punto de vista único. El carácter único se logra debido a que los esquemas de diferencias se

tratan como ecuaciones operacionales u operacionales de diferencias con operadores que actúan en un espacio de dimensión finita dotado de un producto escalar. Al construir la teoría de los métodos iterativos y la de estabilidad de los esquemas de diferencias se emplean las propiedades más simples de los operadores (de las matrices): el carácter constante de los signos, la autoconjugación, ciertas propiedades de los valores propios y de los vectores propios; no se hacen ningunas suposiciones referentes a la estructura de los operadores. Todas las condiciones de la teoría resultaron ser muy cómodas para la comprobación en el caso de esquemas concretos de diferencias. El material expuesto en los cap. VI y VII puede servir para un estudio más completo de la teoría la cual se da en los libros [6, 9].

Capítulo I

Ecuaciones en diferencias

En el presente capítulo se estudian funciones reticulares, cuyo argumento es un número entero, y además ecuaciones en diferencias de segundo orden. Se da a conocer un aparato matemático más simple para el estudio de las funciones reticulares y de los operadores de diferencias. Para resolver las ecuaciones en diferencias de segundo orden se emplea el método de eliminación llamado método de factorización.

§ 1. Funciones reticulares

1. Funciones reticulares y operaciones sobre ellas. Ya se ha mencionado que en los métodos aproximados las funciones de un argumento continuo se sustituyen habitualmente por las de argumento discreto, esto es, por las funciones reticulares. La *función reticular* puede, pues, considerarse como una función cuyo argumento es un número entero:

$$y(i) = y_i, \qquad i = 0, \pm 1, \pm 2, \ldots$$

Podemos introducir para $y(i)$ las operaciones que representan un análogo discreto (de diferencias) de las operaciones de diferenciación e integración.

El análogo de la primera derivada lo constituyen las diferencias de *primer orden*:

$$\Delta y_i = y_{i+1} - y_i, \text{ la diferencia derecha};$$

$$\nabla y_i = y_i - y_{i-1}, \text{ la diferencia izquierda};$$

$$\delta y_i = \frac{1}{2}(\Delta y_i + \nabla y_i) = \frac{1}{2}(y_{i+1} - y_{i-1}), \text{ la diferencia cen-}$$

tral; resulta fácil notar en este caso que $\Delta y_i = \nabla y_{i+1}$.

Ahora podemos escribir las diferencias de *segundo orden*:

$$\Delta^2 y_i = \Delta(\Delta y_i) = \Delta(y_{i+1} - y_i) = y_{i+2} - 2y_{i+1} + y_i,$$

$$\Delta\nabla y_i = \Delta(y_i - y_{i-1}) = (y_{i+1} - y_i) - (y_i - y_{i-1}) =$$
$$= y_{i+1} - 2y_i + y_{i-1},$$

de modo que

$$\Delta^2 y_i = \Delta\nabla y_{i+1}.$$

Análogamente se define la diferencia de *m-ésimo orden*:

$$\Delta^m y_i = \Delta(\Delta^{m-1} y_i),$$

que contiene los valores de $y_i, y_{i+1}, \ldots, y_{i+m}$. Es evidente que

$$\sum_{j=k}^{i} \Delta y_j = y_{i+1} - y_k, \qquad \sum_{j=k}^{i} \nabla y_j = y_i - y_{k-1}.$$

2. Análogos en diferencias de las fórmulas de diferenciación de un producto y de integración por partes. Sean y_i, v_i las funciones arbitrarias cuyo argumento es un número entero. En este caso serán válidas las fórmulas

$$\Delta(y_i v_i) = y_i \Delta v_i + v_{i+1} \Delta y_i = y_{i+1} \Delta v_i + v_i \Delta y_i, \qquad (1)$$

$$\nabla(y_i v_i) = y_{i-1} \nabla v_i + v_i \nabla y_i = y_i \nabla v_i + v_{i-1} \nabla y_i, \qquad (2)$$

que se comprueban inmediatamente. Por ejemplo,

$$\Delta(y_i v_i) = y_{i+1} v_{i+1} - y_i v_i;$$

$$y_i \Delta v_i + v_{i+1} \Delta y_i = y_i(v_{i+1} - v_i) + v_{i+1}(y_{i+1} - y_i) =$$
$$= y_{i+1} v_{i+1} - y_i v_i = \Delta(y_i v_i).$$

Al deducir la fórmula para $\nabla(y_i v_i)$ es suficiente tomar en consideración que $\nabla(y_i v_i) = \Delta(y_{i-1} v_{i-1})$.

Las fórmulas (1), (2) representan los análogos de la fórmula de diferenciación del producto $(y(x)\, v(x))' = yv' + vy'$.

Como análogo de la fórmula de integración por partes interviene la fórmula de sumación por partes:

$$\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i + (yv)_N - (yv)_0, \qquad (3)$$

la cual se anota también en la forma

$$\sum_{i=1}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N-1} v_i \nabla y_i + y_{N-1} v_N - y_0 v_1. \qquad (4)$$

Para deducir la fórmula (3) hagamos uso de la fórmula (1); tenemos

$$y_i \Delta v_i = \Delta (y_i v_i) - v_{i+1} \Delta y_i = \Delta (y_i v_i) - v_{i+1} \nabla y_{i+1},$$

puesto que $\Delta y_i = \nabla y_{i+1}$; de aquí obtenemos

$$\sum_{i=0}^{N-1} y_i \Delta v_i + \sum_{i=1}^{N} v_i \nabla y_i =$$

$$= \sum_{i=0}^{N-1} \Delta (y_i v_i) - \sum_{i=0}^{N-1} v_{i+1} \nabla y_{i+1} + \sum_{i=1}^{N} v_i \nabla y_i =$$

$$= y_N v_N - y_0 v_0 - \sum_{i=1}^{N} v_i \nabla y_i + \sum_{i=1}^{N} v_i \nabla y_i = (yv)_N - (yv)_0.$$

Si $y_0 = 0$, $y_N = 0$, entonces $\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i$.

La fórmula de sumación por partes puede emplearse para calcular sumas.

EJEMPLOS. 1. Calcúlese la suma $S_N = \sum_{i=1}^{N} i 2^i$. Ponemos $v_i = i$, $\nabla y_i = 2^i$, de suerte que

$$y_i = y_{i-1} + 2^i = y_0 + \sum_{j=1}^{i} 2^j = y_0 + 2^{i+1} - 2.$$

Elijamos $y_0 = 2 - 2^{N+1}$; entonces $y_N = 0$. Como $v_0 = 0$ y $\Delta v_i = 1$, de (3) se infiere

$$S_N = \sum_{i=1}^{N} v_i \nabla y_i = -\sum_{i=0}^{N-1} v_i \Delta v_i = -\sum_{i=0}^{N-1} y_i =$$

$$= -N (y_0 - 2) - \sum_{i=0}^{N-1} 2^{i+1} = N 2^{N+1} - (2^{N+1} - 2),$$

de manera que $S_N = (N-1) 2^{N+1} + 2$.

2. Calcúlese $S_N = \sum_{i=1}^{N} i(i-1) = \sum_{i=1}^{N-1} i(i+1)$. Ponemos $y_i = i$, $\nabla v_i = i+1$. En este caso $v_{i+1} = v_i + (i+1) = v_1 + (2+3+\ldots+(i+1)) = (v_1 - 1) + (i+1) \times (i+2)/2$, $v_i = v_1 - 1 + i(i+1)/2$. Elijamos v_1 de la condición $v_N = 0$, es decir, $v_1 = 1 - N(N+1)/2$. Aplicando la fórmula (3) y teniendo en cuenta que $y_0 = 0$, $v_N = 0$, $\nabla y_i = 1$, encontramos

$$S_N = \sum_{i=1}^{N-1} i(i+1) = \sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i =$$
$$= -\sum_{i=1}^{N-1} v_i = -(N-1)(v_1 - 1) - \frac{1}{2} \sum_{i=1}^{N-1} i(i+1) =$$
$$= -\frac{1}{2} S_N + \frac{(N-1)N(N+1)}{2},$$

de modo que $S_N = \frac{1}{3}(N-1)N(N+1)$. De aquí se deduce que

$$\sum_{i=1}^{N} i^2 = 1^2 + 2^2 + \ldots + N^2 = S_N + \sum_{i=1}^{N} i = \frac{N(N+1)(2N+1)}{6}.$$

§ 2. Ecuaciones en diferencias

1. Ecuaciones en diferencias. Una ecuación lineal respecto de la función reticular $y_i = y(i)$ $(i = 0, \pm 1, \pm 2, \ldots)$

$$a_0(i)\, y(i) + a_1(i)\, y(i+1) + \ldots + a_m(i)\, y(i+m) = f(i), \qquad (1)$$

donde $a_k(i)$ $(k = 0, 1, \ldots, m)$, $f(i)$ son las funciones reticulares prefijadas, $a_0(i) \neq 0$, $a_m(i) \neq 0$, lleva el nombre de *ecuación en diferencias lineal* de m-ésimo orden. Dicha ecuación contiene $m+1$ valores de la función $y(i)$.

Haciendo uso de las fórmulas para las diferencias Δy_i, $\Delta^2 y_i, \ldots, \Delta^{m-1} y_i$, podemos expresar los valores y_{i+1}, $y_{i+2}, \ldots, y_{m+1}$ en términos de y_i y las diferencias citadas: $y_{i+1} = y_i + \Delta y_i$, $\quad y_{i+2} = \Delta^2 y_i + 2y_{i+1} - y_i = \Delta^2 y_i +$

$+ 2\Delta y_i + y_i$, etc. Como resultado, obtendremos de (1) una nueva notación de la *ecuación en diferencias* de m-ésimo orden:

$$\alpha_0(i)\,y_i + \alpha_1(i)\,\Delta y_i + \ldots + \alpha_m(i)\,\Delta^m y_i = f(i),$$
$$i = 0, \pm 1, \pm 2, \ldots, \quad (2)$$

(por lo que se determina precisamente el término «ecuación en o de diferencias»). Si los coeficientes $a_0, a_1, \ldots, a_m$ no dependen de i, $a_0 \neq 0$ y $a_m \neq 0$, entonces (1) se denomina ecuación en diferencias lineal de m-ésimo orden con *coeficientes constantes*.

Para $m = 1$ de (1) se obtiene una ecuación en diferencias de *primer orden*

$$a_0(i)\,y_i + a_1(i)\,y_{i+1} = f(i), \qquad a_0(i) \neq 0, \qquad a_1(i) \neq 0; \quad (3)$$

cuando $m = 2$, obtenemos una ecuación en diferencias de *segundo orden*

$$a_0(i)\,y_i + a_1(i)\,y_{i+1} + a_2(i)\,y_{i+2} = f(i), \qquad a_0(i) \neq 0, \quad a_2(i) \neq 0.$$

Nos limitaremos al estudio de las ecuaciones en diferencias de primero y segundo órdenes.

2. Ecuaciones de primer orden. Examinemos la ecuación en diferencias de primer orden (3). Al sustituir $y_{i+1} = y_i + \Delta y_i$, obtendremos

$$\bar{a}_0(i)\,y_i + a_1(i)\,\Delta y_i = f(i), \qquad \bar{a}_0 = a_0 + a_1.$$

Como ejemplos más simples de ecuaciones en diferencias de primer orden pueden servir las ecuaciones para los términos de una progresión aritmética $y_{i+1} = y_i + d$ y de una progresión geométrica $y_{i+1} = qy_i$.

Escribamos la ecuación (3) en la forma

$$y_{i+1} = q_i y_i + \varphi_i, \quad (4)$$

donde $q_i = -a_0(i)/a_1(i)$, $\varphi_i = f(i)/a_1(i)$. De aquí se ve que la solución $y(i)$ está definida unívocamente para $i > i_0$, si está prefijado el valor $y(i_0)$. Supongamos que para $i = 0$ viene prefijado $y_0 = y(0)$. En tal caso podemos determinar

$y_1, y_2, \ldots, y_i, \ldots$ Eliminando sucesivamente según la fórmula (4) $y_i, y_{i-1}, \ldots, y_1$, obtendremos

$$y_{i+1} = q_i q_{i-1} \ldots q_0 y_0 + \varphi_i + q_i \varphi_{i-1} + \\ + q_i q_{i-1} \varphi_{i-2} + \ldots + q_i q_{i-1} \ldots q_1 \varphi_0,$$

o bien

$$y_{i+1} = \left(\prod_{k=0}^{i} q_k\right) y_0 + \sum_{k=0}^{i-1} \left(\prod_{s=k+1}^{i} q_s\right) \varphi_k + \varphi_i. \qquad (5)$$

Para la ecuación con coeficientes constantes $q_i = q$, de lo que se tiene

$$y_{i+1} = q^{i+1} y_0 + \sum_{k=0}^{i} q^{i-k} \varphi_k, \qquad i = 0, 1, 2, \ldots, \qquad (6)$$

que es una solución de la ecuación en diferencias (4) con coeficientes constantes.

3. Desigualdades de primer orden. Si el signo de igualdad en las expresiones de tipo (1) ó (2) lo sustituimos por los signos de desigualdad $<$, $>$, $\leqslant$, $\geqslant$, obtendremos *desigualdades en diferencias* de m-ésimo orden. Sea dada una desigualdad en diferencias de primer orden

$$y_{i+1} \leqslant q y_i + f_i, \qquad i = 0, 1, 2, \ldots, q \geqslant 0; \qquad (7)$$

sin restringir la generalidad de nuestros razonamientos, en adelante consideramos siempre que $q > 0$ (y_0, q, f_i son conocidos). Hallemos la solución de la desigualdad citada. Sea v_i una solución de la ecuación en diferencias

$$v_{i+1} = q v_i + f_i, \qquad i = 0, 1, \ldots, \qquad v_0 = y_0. \qquad (8)$$

En este caso queda lícita la estimación

$$y_i \leqslant v_i. \qquad (9)$$

En efecto, al sustraer (8) de (7), encontramos

$$y_{i+1} - v_{i+1} \leqslant q (y_i - v_i) \leqslant q^2 (y_{i-1} - v_{i-1}) \leqslant \ldots \\ \ldots \leqslant q^{i+1} (y_0 - v_0) = 0.$$

Al poner en (9) la expresión explícita para v_i, tenemos

$$y_i \leqslant q^i y_0 + \sum_{k=0}^{i-1} q^{i-1-k} f_k, \qquad i = 0, 1, 2, \ldots, \qquad (10)$$

lo que es la solución de la desigualdad (7).

4. Ecuaciones de segundo orden con coeficientes constantes. Analicemos una ecuación en diferencias de segundo orden

$$by_{i+1} - cy_i + ay_{i-1} = f_i, \qquad i = 0, 1, \ldots, \qquad a \neq 0, \qquad b \neq 0, \tag{11}$$

cuyos coeficientes no dependen de i. Si $f_i = 0$, la ecuación

$$by_{i+1} - cy_i + ay_{i-1} = 0, \qquad i = 0, 1, \ldots, \tag{12}$$

se llamará *homogénea*. Su solución se halla en la forma explícita.

Sea $\bar{y}_i$ una solución de la ecuación homogénea (12), y sea y_i^* una solución cualquiera de la ecuación no homogénea (11). Entonces, la suma $y_i = \bar{y}_i + y_i^*$ será también una solución de la ecuación no homogénea:

$$b(\bar{y}_{i+1} + y_i^*) - c(\bar{y}_i + y_i^*) + a(\bar{y}_{i-1} + y_{i-1}) =$$
$$= [b\bar{y}_{i+1} - c\bar{y}_i + a\bar{y}_{i-1}] + [by_{i+1}^* - cy_i^* + ay_{i-1}^*] = f_i.$$

Esta propiedad se debe a la linealidad de la ecuación (11); ella queda en vigor para la ecuación en diferencias (1) de cualquier orden. Es evidente que si $\bar{y}_i$ es una solución de la ecuación homogénea (12), entonces también $c\bar{y}_i$ (donde c es una constante arbitraria) satisface la citada ecuación.

Sean $y_i^{(1)}$ e $y_i^{(2)}$ dos soluciones de la ecuación (12). Se denominarán *linealmente independientes*, si la igualdad

$$c_1 y_i^{(1)} + c_2 y_i^{(2)} = 0, \quad i = 0, 1, 2, \ldots,$$

se verifica sólo cuando $c_1 = c_2 = 0$. Esta afirmación es equivalente a la exigencia de que el determinante del sistema

$$c_1 y_i^{(1)} + c_2 y_i^{(2)} = 0,$$
$$c_1 y_{i+m}^{(1)} + c_2 y_{i+m}^{(2)} = 0, \qquad m = \pm 1, \pm 2, \ldots,$$

sea distinto de cero para cualesquiera i, m. En particular,

$$\Delta_{i,\, i+1} = \begin{vmatrix} y_i^{(1)} & y_i^{(2)} \\ y_{i+1}^{(1)} & y_{i+1}^{(2)} \end{vmatrix} \neq 0.$$

Al igual que en la teoría de las ecuaciones diferenciales, se puede intro cir la noción de *solución general* de la

ecuación en diferencias (12) y mostrar que si las soluciones $y_i^{(1)}$, $y_i^{(2)}$ son linealmente independientes, la solución general de la ecuación (12) tendrá la forma

$$y_i = c_1 y_i^{(1)} + c_2 y_i^{(2)},$$

donde c_1 y c_2 son unas constantes arbitrarias. La solución general de la ecuación no homogénea (11) puede representarse en la forma

$$y_i = c_1 y_i^{(1)} + c_2 y_i^{(2)} + y_i^*, \tag{13}$$

donde y_i^* es una solución (particular) cualquiera de la ecuación (11). Lo mismo que en el caso de las ecuaciones diferenciales, para determinar c_1 y c_2 se deben definir las condiciones complementarias iniciales o las de contorno.

La solución particular de la ecuación (12) puede hallarse en la forma explícita. Buscaremos dicha solución en la forma $y_i = q^i$, donde $q \neq 0$ es un número hasta ahora desconocido. Al realizar la sustitución $y_k = q^k$ en (12), obtendremos una ecuación cuadrática $bq^2 - cq + a = 0$, cuyas raíces son

$$q_1 = \frac{c + \sqrt{c^2 - 4ab}}{2b}, \quad q_2 = \frac{c - \sqrt{c^2 - 4ab}}{2b}. \tag{14}$$

Según sean los valores del discriminante $D = c^2 - 4ab$, son posibles tres casos:

1) $D = c^2 - 4ab > 0$. Las raíces q_1 y q_2 son reales y distintas. Les corresponden las soluciones particulares

$$y_k^{(1)} = q_1^k, \qquad y_k^{(2)} = q_2^k;$$

estas soluciones son linealmente independientes, puesto que es distinto de cero el determinante:

$$\Delta_{k,\,k+1} = \begin{vmatrix} q_1^k & q_1^{k+1} \\ q_2^k & q_2^{k+1} \end{vmatrix} = q_1^k q_2^k (q_2 - q_1) \neq 0$$

Ha de notarse que $q_1 \neq 0$ y $q_2 \neq 0$, pues en el caso contrario $a = 0$ y la ecuación (12) no sería ecuación en diferencias de segundo orden. La solución general de la ecuación (12) tiene por expresión

$$y_k = c_1 q_1^k + c_2 q_2^k. \tag{15}$$

2) $D = c^2 - 4ab < 0$. La ecuación cuadrática cuenta con las raíces complejas conjugadas

$$q_1 = \frac{c + i\sqrt{|D|}}{2b}; \qquad q_2 = \frac{c - i\sqrt{|D|}}{2b},$$

donde i es la unidad imaginaria. Resulta cómodo representar estas raíces en la forma

$$q_1 = \rho e^{i\varphi}, \qquad q_2 = \rho e^{-i\varphi}, \qquad \rho = \sqrt{\frac{a}{b}},$$

$$\varphi = \operatorname{arctg} \frac{\sqrt{|D|}}{c}.$$

No sólo las funciones

$$q_1^k = \rho^k e^{ik\varphi} = \rho^k (\cos k\varphi - i \operatorname{sen} k\varphi),$$

$$q_2^k = \rho^k e^{-ik\varphi} = \rho^k (\cos k\varphi - i \operatorname{sen} k\varphi)$$

representan soluciones particulares sino también las funciones siguientes:

$$y_k^{(1)} = \rho^k \cos k\varphi, \qquad y_k^{(2)} = \rho^k \operatorname{sen} k\varphi,$$

las cuales son linealmente independientes en virtud de la independencia lineal de las funciones sen $k\varphi$ y cos $k\varphi$. La solución general tiene la forma

$$y_k = \rho^k (c_1 \cos k\varphi + c_2 \operatorname{sen} k\varphi). \tag{16}$$

3) $D = c^2 - 4ab = 0$. Las raíces son reales e iguales: $q_1 = q_2 = c/(2b) = q_0$. Las soluciones

$$y_k^{(1)} = q_0^k, \qquad y_k^{(2)} = kq_0^k \tag{17}$$

son linealmente independientes. Mostraremos que $y_k^{(2)}$ es una solución de la ecuación (12):

$$by_{k+1}^{(2)} - cy_k^{(2)} + ay_{k-1}^{(2)} = b(k+1)\,q_0^{k+1} - ckq_0^k + a(k-1)\,q_0^{k-1} =$$
$$= k\,(bq_0^{k+1} - cq_0^k + aq_0^{k-1}) + (bq_0^2 - a)\,q_0^{k-1} = 0,$$

puesto que $bq_0^2 - a = b\frac{c^2}{4b^2} - a = \frac{D}{4b} = 0$. Como $\Delta_{k,\,k+1} = \begin{vmatrix} q_0^k & kq_0^k \\ q_0^{k+1} & (k+1)\,q_0^{k+1} \end{vmatrix} = q_0^{2k+1} \neq 0$, entonces las solu-

ciones (17) son linealmente independientes y la solución general tiene por expresión

$$y_k = c_1 q_0^k + c_2 k q_0^k.$$

5. Ejemplos. Veamos algunos ejemplos de resolución de las ecuaciones en diferencias de segundo orden (11).

1. Hállese la solución general de la ecuación

$$y_{k+1} - 2py_k + y_{k-1} = 0, \qquad a = b = 1, \qquad c = 2p > 0.$$

Son posibles tres casos. 1) $p < 1$. Ponemos $p = \cos\alpha$; entonces $D = 4(\cos^2\alpha - 1) = -4\operatorname{sen}^2\alpha < 0$. Las soluciones particulares tienen la forma

$$y_k^{(1)} = \cos k\alpha, \qquad y_k^{(2)} = \operatorname{sen} k\alpha.$$

2) $p > 1$. Suponiendo $p = \operatorname{ch}\alpha$, obtendremos para q una ecuación cuadrática $q^2 - 2\operatorname{ch}\alpha q + 1 = 0$; su discriminante es $D = 4(\operatorname{ch}^2\alpha - 1) = 4\operatorname{sh}^2\alpha$, mientras que las raíces tienen por expresión $q_{1,2} = \operatorname{ch}\alpha \pm \operatorname{sh}\alpha = e^{\pm\alpha}$. El papel de soluciones particulares desempeñan las funciones

$$y_k^{(1)} = \operatorname{ch} k\alpha, \qquad y_k^{(2)} = \operatorname{sh} k\alpha.$$

3) $p = 1$. En este caso $q^2 - 2q + 1 = 0$, $q_{1,2} = 1$, las soluciones particulares tienen la forma $y_k^{(1)} = 1$, $y_k^{(2)} = k$, y la solución general es

$$y_k = c_1 + c_2 k.$$

2. Hállese la solución de la ecuación

$$y_{k+2} - y_{k+1} = 2y_k = 0.$$

El discriminante es igual a $D = 1 + 8 = 9$, las raíces serán $q_{1,2} = (1 \pm 3)/2$, $q_1 = 2$, $q_2 = -1$. La solución general es de la forma

$$y_k = c_1 2^k + c_2(-1)^k.$$

3. Hállese la solución general de la ecuación

$$y_{k+1} - y_k - 6y_{k-1} = 2^{k+1}. \tag{18}$$

La solución general de una ecuación no homogénea es la suma $y_k = \overline{y}_k + y_k^*$ de la solución general $\overline{y}_k$ de la ecuación homogénea y de la solución particular y_k^* de la ecuación

no homogénea. Hallemos primero la solución general de la ecuación homogénea. El discriminante es $D = 1 + 24 = 25 > 0$, y las raíces de la ecuación cuadrática $q^2 - q - 6 = 0$ son $q_1 = 3$, $q_2 = -2$, de suerte que $y_k^{(1)} = 3^k$, $y_k^{(2)} = (-2)^k$. La solución particular y_k^* buscaremos en la forma $y_k^* = c2^k$, donde $c = \text{const}$. Sustituyendo $y_k^* = c2^k$ en (18), obtendremos $c(2^{k+1} - 2^k - 6\cdot 2^{k-1}) = c\cdot 2^{k-1}(-4) = 2^{k+1}$, $c = -1$.

La solución general de la ecuación (18) tiene por expresión

$$y_k = c_1\cdot 3^k + c_2(-2)^k - 2^k.$$

6. Ecuación en diferencias de segundo orden con coeficientes variables. Problema de Cauchy y problema de contorno. Examinemos ahora una ecuación en diferencias con coeficientes variables

$$b_i y_{i+1} - c_i y_i + a_i y_{i-1} = f_i,$$
$$a_i \neq 0, \quad b_i \neq 0, \quad i = 0, 1, 2, \dots \qquad (19)$$

Dado que $b_i \neq 0$, de (19) obtenemos la siguiente relación recurrente:

$$y_{i+1} = \frac{c_i y_i - a_i y_{i-1} + f_i}{b_i}, \qquad b_i \neq 0. \qquad (20)$$

Expresemos y_{i+1} e y_{i-1} en términos de y_i y las diferencias de primero y segundo órdenes. La ecuación (19) se anotará en este caso en la forma

$$\Delta\nabla y_i + (b_i - a_i)\,\Delta y_i - (c_i - a_i - b_i)\,y_i = f_i,$$
$$a_i \neq 0, \qquad b_i \neq 0.$$

La solución de una ecuación en diferencias de primer orden depende de una constante arbitraria y se determina unívocamente, si está prefijada una condición complementaria, por ejemplo, $y_0 = c_0$. La solución de la ecuación de segundo orden se determina por dos constantes arbitrarias y se puede hallarla, siempre que vienen dadas dos condiciones complementarias. Si ambas condiciones están dadas en dos puntos vecinos, se trata de un *problema de Cauchy*. Si las dos condiciones están dadas en dos puntos diferentes (pero no vecinos), obtenemos un *problema de contorno*. De mayor interés para nosotros serán precisamente los

problemas de contorno. Introduzcamos las designaciones

$$Ly_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1}$$

y enunciemos los problemas mencionados más detalladamente.

PROBLEMA DE CAUCHY: hállese la solución de la ecuación

$$Ly_i = f_i, \qquad i = 1, 2, \ldots, \tag{21}$$

con las condiciones complementarias

$$y_0 = \mu_1, \qquad y_1 = \mu_2. \tag{22}$$

La segunda condición de (22) puede notarse de otro modo: $\Delta y_0 = y_1 - y_0 = \mu_2 - \mu_1 = \bar{\mu}_1$; podemos, entonces, decir que en el caso del problema de Cauchy vienen dadas en un punto $i = 0$ las magnitudes

$$y_0 = \mu_1, \qquad \Delta y_0 = \bar{\mu}_1. \tag{22'}$$

PROBLEMA DE CONTORNO: hállese la solución de la ecuación

$$Ly_i = f_i, \qquad i = 1, 2, \ldots, N-1$$

para las condiciones complementarias

$$y_0 = \mu_1, \qquad y_N = \mu_2, \qquad N \geqslant 2. \tag{23}$$

En los nodos de frontera $i = 0$ e $i = N$ pueden definirse no sólo los valores de las funciones, sino también sus diferencias y combinaciones, es decir, las expresiones $\alpha_1 \Delta y_0 + \beta_1 y_0$ para $i = 0$, y $\alpha_2 \nabla y_N + \beta_2 y_N$ para $i = N$. Tales condiciones se pueden anotar en la forma

$$y_0 = \varkappa_1 y_1 + \mu_1, \qquad y_N = \varkappa_2 y_{N-1} + \mu_2. \tag{24}$$

Si $\varkappa_1 = \varkappa_2 = 0$, obtenemos de aquí las *condiciones de primer género*; cuando $\varkappa_1 = 1$, $\varkappa_2 = 1$, tenemos *condiciones de segundo género*

$$\Delta y_0 = -\mu_1, \qquad \nabla y_N = \mu_2. \tag{25}$$

Si $\varkappa_{1,2} \neq 0; 1$, (24) llevan el nombre de *condiciones de tercer género*:

$$-\varkappa_1 \Delta y_0 + (1 - \varkappa_1)\, y_0 = \mu_1,$$
$$\varkappa_2 \nabla y_N + (1 - \varkappa_2)\, y_N = \mu_2. \tag{26}$$

Además, pueden encontrarse problemas de contorno con ciertas combinaciones de las citadas condiciones de con-

torno: las condiciones de un tipo para $i = 0$, y condiciones de otro tipo, para $i = N$.

La solución del problema de Cauchy se halla directamente de la ecuación (21) según la fórmula recurrente (20), tomando en consideración los datos iniciales $y_0 = \mu_1$, $y_1 = \mu_2$. Para la resolución de los problemas de contorno se emplea un método más complejo (método de eliminaciones) el cual se expondrá en adelante.

Para una ecuación con coeficientes constantes la solución del problema de contorno puede expresarse en la forma explícita.

EJEMPLO. Hállese la solución del problema de contorno

$$\Delta^2 y_{i-1} = 1, \quad i = 1, 2, \ldots, N-1, \quad y_0 = 0, \quad y_N = 0. \tag{27}$$

La ecuación homogénea $\Delta^2 y_{i-1} = y_{i+1} - 2y_i + y_{i-1} = 0$ tiene su solución general $\overline{y}_i = c_1 + c_2 i$. La solución particular y_i^* de la ecuación no homogénea $\Delta^2 y_{i-1} = y_{i+1} - 2y_i + y_{i-1} = 1$ se buscará en la forma $y_i^* = ci^2$. Al sustituir esta expresión en la ecuación (27), encontramos $\Delta^2 y_{i-1}^* = c((i+1)^2 - 2i^2 + (i-1)^2) = 1$, es decir, $c = 1/2$, de suerte que $y_i = \overline{y}_i + y_i^* = c_1 + c_2 i + i^2/2$. Con el fin de determinar c_1 y c_2 se usan las condiciones de contorno para $i = 0$, $i = N$: $y_0 = c_1 = 0$, $y_N = c_2 N + N^2/2 = 0$, $c_2 = -N/2$. De este modo,

$$y_i = -\frac{1}{2} iN + \frac{1}{2} i^2 = -\frac{1}{2} i(N-i)$$

es la solución del problema (27).

§ 3. Resolución de los problemas de contorno en diferencias para las ecuaciones de segundo orden

1. Resolución de los problemas de contorno en diferencias por el método de factorización. Un problema de contorno

$$a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -f_i, \quad a_i \neq 0, \quad b_i \neq 0,$$
$$i = 1, 2, \ldots, N-1, \tag{1}$$
$$y_0 = \varkappa_1 y_1 + \mu_1, \qquad y_N = \varkappa_2 y_{N-1} + \mu_2$$

representa el sistema de ecuaciones algebraicas lineales con matriz tridiagonal de dimensión $(N+1)\times(N+1)$:

$$A=\begin{bmatrix} 1 & -\varkappa_1 & 0 \dots 0 & 0 & 0 \dots 0 & 0 & 0 \\ a_1 & -c_1 & b_1 \dots 0 & 0 & 0 \dots 0 \dots & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 \dots a_i & -c_i & b_i \dots 0 & 0 & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & 0 \dots 0 & 0 & 0 \dots a_{N-1} & -c_{N-1} & b_{N-1} \\ 0 & 0 & 0 \dots 0 & 0 & 0 \dots 0 & -\varkappa_2 & 1 \end{bmatrix}.$$

En lugar de (1) podemos escribir

$$Ay=f,\quad y=(y_0, y_1, \dots, y_N),$$
$$f=(\mu_1, -f_1, \dots, -f_{N-1}, \mu_2). \qquad (2)$$

En el caso del primer problema del contorno la matriz correspondiente tiene la dimensión $(N-1)\times(N-1)$.

Para resolver el problema de contorno (1) puede emplearse el siguiente método de eliminación llamado *método de factorización*. Supongamos que tiene lugar la relación

$$y_i=\alpha_{i+1}y_{i+1}+\beta_{i+1} \qquad (3)$$

con coeficientes indeterminados α_{i+1} y β_{i+1}, y sustituyamos $y_{i-1}=\alpha_i y_i+\beta_i$ en (1):

$$(a_i\alpha_i-c_i)\,y_i+b_iy_{i+1}=-(f_i+a_i\beta_i);$$

al comparar esta identidad con (3), encontramos

$$\alpha_{i+1}=\frac{b_i}{c_i-a_i\alpha_i},\quad i=1,\ 2,\ \dots,\ N-1, \qquad (4)$$

$$\beta_{i+1}=\frac{a_i\beta_i+f_i}{c_i-a_i\alpha_i},\quad i=1,\ 2,\ \dots,\ N-1. \qquad (5)$$

Con el fin de hallar α_1, β_1 hagamos uso de la condición de contorno para $i=0$. De las fórmulas (3) y (1) encontramos para $i=0$:

$$\alpha_1=\varkappa_1,\qquad \beta_1=\mu_1. \qquad (6)$$

Conociendo α_1, β_1 y pasando de i a $i+1$ en las fórmulas (4) y (5), determinamos α_i y β_i para cualquier $i=2,\ 3,\ \dots$ $\dots,\ N$. Los cálculos según la fórmula (3) se llevan a cabo

pasando de $i+1$ a i (es decir, conociendo y_{i+1}, hallamos y_i), y para iniciar los cálculos mencionados se debe prefijar y_N. Determinemos y_N partiendo de la condición de contorno $y_N = \varkappa_2 y_{N-1} + \mu_2$ y de la condición (3) para $i = N-1$: $y_{N-1} = \alpha_N y_N + \beta_N$. De aquí encontramos

$$y_N = \frac{\mu_2 + \varkappa_2 \beta_N}{1 - \alpha_N \varkappa_2}. \qquad (7)$$

Reunamos ahora todas las fórmulas de factorización y escribámoslas en el orden de su aplicación:

$$\overset{(\rightarrow)}{\alpha_{i+1}} = \frac{b_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \ldots, N-1, \quad \alpha_1 = \varkappa_1; \qquad (8)$$

$$\overset{(\rightarrow)}{\beta_{i+1}} = \frac{a_i \beta_i + f_i}{c_i - a_i \alpha_i}, \quad i = 1, 2, \ldots, N-1, \quad \beta_1 = \mu_1; \qquad (9)$$

$$\overset{(\leftarrow)}{y_i} = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad i = N-1, N-2, \ldots, 2, 1, 0,$$

$$y_N = \frac{\mu_2 + \varkappa_2 \beta_N}{1 - \alpha_N \varkappa_2}. \qquad (10)$$

Las flechas indican la dirección de cálculo: $(\rightarrow)$ de i a $i+1$, $(\leftarrow)$ de $i+1$ a i.

De este modo, el problema de contorno para la ecuación de segundo orden se ha reducido a tres problemas de Cauchy para las ecuaciones de primer orden.

2. Estabilidad del método de factorización. Las fórmulas de factorización pueden emplearse, si los denominadores de las fracciones (8) y (10) no se reducen a cero. Las condiciones suficientes para ello están representadas por las desigualdades

$$|c_i| \geqslant |a_i| + |b_i|, \qquad i = 1, 2, \ldots, N-1,$$

$$|\varkappa_1| \leqslant 1, \qquad |\varkappa_2| \leqslant 1, \qquad |\varkappa_1| + |\varkappa_2| < 2. \quad (11)$$

Probemos que siendo cumplidas las condiciones (11), los denominadores $c_i - a_i \alpha_i$ y $1 - \alpha_N \varkappa_2$ no se reducen a cero y que

$$|\alpha_i| \leqslant 1, \qquad i = 1, 2, \ldots, N. \qquad (12)$$

Supongamos que $|\alpha_i| \leqslant 1$, y mostremos que $|\alpha_{i+1}| \leqslant 1$; entonces de aquí y de la condición $|\alpha_1| = |\varkappa_1| \leqslant 1$ se deducirá (12). Examinaremos la diferencia $|c_i - a_i \alpha_i| -$

$- |b_i| \geqslant |c_i| - |a_i||\alpha_i| - |b_i| \geqslant |a_i|(1 - |\alpha_i|) \geqslant$ $\geqslant 0$, de modo que $|c_i - a_i\alpha_i| \geqslant |b_i| > 0$, y $|\alpha_{i+1}| =$ $= |b_i|/|c_i - a_i\alpha_i| \leqslant 1$.

Observemos que si $|c_{i_0}| > |a_{i_0}| + |b_{i_0}|$ siquiera en un solo punto $i = i_0$, entonces $|\alpha_i| < 1$ para todo $i > i_0$, incluso para $i = N$: $|\alpha_N| < 1$. En este caso $|1 - \alpha_N\varkappa_2| \geqslant$ $\geqslant 1 - |\alpha_N||\varkappa_2| \geqslant 1 - |\alpha_N| > 0$, y la condición $|\varkappa_1| + |\varkappa_2| < 2$ será superflua. Si $|\varkappa_1| < 1$, entonces $|\alpha_N| < 1$. En cambio, si $|\varkappa_1| = 1$, entonces $|\varkappa_2| < 1$ y $|\alpha_N| \leqslant 1$, y tenemos $|1 - \alpha_N\varkappa_2| \geqslant 1 - |\alpha_N||\varkappa_2| \geqslant$ $\geqslant 1 - |\varkappa_2| > 0$. De este modo, si se cumplen las condiciones (11), el problema (1) tiene la única solución la cual se halla según las fórmulas de factorización (8)—(10).

Los cálculos según las fórmulas (8)—(10) se realizan en un ordenador aproximadamente, con un número finito de cifras significativas. A consecuencia de los errores de redondeo se halla, de hecho, no la función y_i (la cual representa solución del problema (1)), sino $\tilde{y}_i$, esto es, la solución del mismo problema con coeficientes perturbados $\tilde{a}_i$, $\tilde{b}_i$, $\tilde{c}_i$, $\tilde{\varkappa}_1$, $\tilde{\varkappa}_2$ y segundos miembros $\tilde{f}_i$, $\tilde{\mu}_1$, $\tilde{\mu}_2$. Surge naturalmente la cuestión de si ocurre o no, en transcurso de los cálculos, el aumento del error de redondeo, lo que puede conducir tanto a la pérdida de precisión, como a la imposibilidad de continuar los cálculos a causa del crecimiento de las magnitudes que se determinan. A título de ejemplo puede servir la búsqueda de y_i según la fórmula $y_{i+1} = qy_i$ para $q > 1$. Puesto que $y_n = q^n y_0$, para cualquier y_0 puede indicarse tal n_0 que y_{n_0} será el infinito de ordenador. Realmente, en virtud de los errores de redondeo, se determina no el valor exacto de y_i, sino el valor de $\tilde{y}_i$ a partir de la ecuación $\tilde{y}_{i+1} = q\tilde{y}_i + \eta$, donde η es el error de redondeo. Para el error $\delta y_i = \tilde{y}_i - y_i$ obtendremos una ecuación $\delta y_{i+1} =$ $= q\delta y_i + \eta$ ($i = 0, 1, \ldots, \delta y_0 = \eta$). De la fórmula $\delta y_i =$ $= q^i\eta + \eta(q^i - 1)/(q - 1)$ se ve que el error δy_i va creciendo, para $q > 1$, de manera exponencial a medida que crece i.

Volvamos al método de factorización y probemos que el error δy_i no aumenta cuando $|\alpha_i| \leqslant 1$. Efectivamente, de

las igualdades $\tilde{y}_i = \alpha_{i+1}\tilde{y}_{i+1} + \beta_{i+1}$, $y_i = \alpha_{i+1}y_{i+1} + \beta_{i+1}$ proviene $\delta y_i = \alpha_{i+1}\delta y_{i+1}$, $|\delta y_i| \leqslant |\alpha_{i+1}|\,|\delta y_{i+1}| \leqslant \leqslant |\delta y_{i+1}|$, porque $|\alpha_{i+1}| \leqslant 1$.

Tomando en consideración que en el transcurso de los cálculos se perturban también los coeficientes α_{i+1}, β_{i+1}, se puede señalar que el error δy_i es proporcional al cuadrado del número de nodos N:

$$\max_{1 \leqslant i \leqslant N} |\delta y_i| \leqslant \varepsilon_0 N^2,$$

donde ε_0 es el error de redondeo. De aquí se ve la relación que existe entre la precisión requerida ε de la solución del problema, el número N de ecuaciones y el número de cifras significativas del ordenador, puesto que $\varepsilon_0 N^2 \approx \varepsilon$.

3. Otras variantes del método de factorización. El método de factorización (8)—(10) analizado más arriba, en el cual la determinación de y_i se realiza sucesivamente de derecha a izquierda, se denomina *factorización derecha*. Análogamente se anotan las fórmulas de la *factorización izquierda*:

$$\overset{(\leftarrow)}{\xi}_i = \frac{a_i}{c_i - b_i\xi_{i+1}}, \quad i = N-1,\ N-2,\ \ldots,\ 2,\ 1, \quad \xi_N = \varkappa_2, \qquad (13)$$

$$\overset{(\leftarrow)}{\eta}_i = \frac{b_i\eta_{i+1} + f_i}{c_i - b_i\xi_{i+1}}, \quad i = N-1,\ N-2,\ \ldots,\ 2,\ 1,\ \eta_N = \mu_2, \qquad (14)$$

$$\overset{(\rightarrow)}{y}_{i+1} = \xi_{i+1}y_i + \eta_{i+1}, \quad i = 0,\ 1,\ \ldots,\ N-1,$$

$$y_0 = \frac{\mu_1 + \varkappa_1\eta_1}{1 - \xi_1\varkappa_1}. \qquad (15)$$

En efecto, suponiendo que $y_{i+1} = \xi_{i+1}y_i + \eta_{i+1}$, eliminemos de (1) y_{i+1}; obtendremos

$$-f_i = a_i y_{i-1} + (b_i\xi_{i+1} - c_i)\, y_i + b_i\eta_{i+1},$$

o bien 1258

$$y = \frac{a_i}{c_i - b_i\xi_{i+1}}\, y_{i-1} + \frac{f_i + b_i\eta_{i+1}}{c_i - b_i\xi_{i+1}}.$$

Al cotejar con la fórmula $y_i = \xi_i y_{i-1} + \beta_i$, obtendremos (13) y (14). El valor de y_0 hallamos de la condición $y_0 = \varkappa_1 y_1 + \mu_1$ y de la fórmula $y_0 = \xi_1 y_1 + \eta_1$. De la desi-

gualdad $|c_i - b_i\xi_{i+1}| \geqslant |c_i| - |b_i|\,|\xi_{i+1}| \geqslant |a_i| + |b_i|(1 - |\xi_{i+1}|)$, $|1 - \xi_1\varkappa_1| \geqslant 1 - |\xi_1|\,|\varkappa_1|$ se ve que las condiciones (11) garantizan que las fórmulas de factorización izquierda sean aplicables y su cálculo sea estable, puesto que $|\xi_i| \leqslant 1$ $(i = 1, 2, \ldots, N)$.

La combinación de las factorizaciones izquierda y derecha da el *método de factorizaciones opuestas*. Empleándose este método, en la región $0 \leqslant i \leqslant i_0 + 1$ se calculan, según las fórmulas (8), (9), los coeficientes de factorización α_i, β_i, y en la región $i_0 \leqslant i \leqslant N$ se hallan, por las fórmulas (13), (14), ξ_i y η_i. Cuando $i = i_0$, se realiza la conjugación de soluciones en la forma (10) y (15).

De las fórmulas $y_{i_0} = \alpha_{i_0+1}y_{i_0+1} + \beta_{i_0+1}y_{i_0+1} = \xi_{i_0+1}y_{i_0} + \eta_{i_0+1}$ hallamos

$$y_{i_0} = \frac{\beta_{i_0+1} + \alpha_{i_0+1}\eta_{i_0}}{1 - \alpha_{i_0+1}\xi_{i_0+1}}.$$

La citada fórmula tiene sentido, puesto que por lo menos una de las magnitudes $|\xi_{i_0+1}|$ ó $|\alpha_{i_0+1}|$ es, en virtud de (11), inferior a la unidad y, por lo tanto, $1 - \alpha_{i_0+1}\xi_{i_0+1} > 0$. Al conocer y_{i_0}, podemos hallar, sirviéndonos de la fórmula (10), todos los valores de y_1 para $i < i_0$, y, por la fórmula (15), todos los valores de y_i para $i > i_0$. Cuando $i > i_0$ e $i < i_0$, los cálculos son autónomos (se llevan a cabo paralelamente). El método de factorizaciones opuestas es particularmente cómodo, si, por ejemplo, se pide hallar y_i sólo en un nodo $i = i_0$.

§ 4. Ecuaciones en diferencias como ecuaciones operacionales

1. Espacio lineal*). Veamos un conjunto H de elementos $x, y, z, \ldots$, respecto de los cuales se sabe que a cada par de elementos x e y, pertenecientes a H, se le pone en correspondencia de tal o cual modo un elemento tercero $z \in H$, llamado suma de los dos elementos primeros y designado $z = x + y$: a todo elemento $x \in H$ y a cada número λ

*) Véase, por ejemplo, V. Ilyin, E. Poznyak, "Linear algebra". Editorial Mir, Moscú, 1985.

se les pone en correspondencia un elemento $u \in H$, denominado producto de x por el número λ y designado $u = \lambda x$.

El conjunto H se llamará *espacio lineal*, si las operaciones de sumación y multiplicación por un número, determinadas para sus elementos $x, y, z, \ldots$, satisfacen los siguientes axiomas:

1) $x + y = y + x$ para cualesquiera $x, y \in H$ (conmutatividad de la sumación);

2) $(x + y) + z = x + (y + z)$ para cualesquiera $x, y, z \in H$ (asociatividad de la sumación);

3) existe un elemento «cero», designado 0, tal que $x + 0 = x$ para cualquier $x \in H$;

4) para todo elemento $x \in H$ existe un elemento opuesto $(-x)$ tal que $x + (-x) = 0$;

5) $1 \cdot x = x$;

6) $(\lambda\mu) x = \lambda (\mu x)$ (asociatividad de la multiplicación);

7) $\lambda (x + y) = \lambda x + \lambda y$; $(\lambda + \mu) x = \lambda x + \mu x$ (distributividad de la multiplicación respecto a sumación), donde λ y μ son unos números cualesquiera.

Un espacio lineal se denomina *complejo*, si para sus elementos está definida la multiplicación por números complejos y se llama *real*, si viene definida solamente la multiplicación por números reales.

Los elementos $x, y, z, \ldots$ del espacio lineal H llevan el nombre de *vectores*.

Los vectores $x_1, x_2, \ldots, x_N$ se denominan *linealmente independientes*, siempre que la igualdad

$$c_1x_1 + c_2x_2 + \ldots + c_Nx_N = 0 \tag{1}$$

se verifica sólo cuando $c_1 = c_2 = \ldots = c_N = 0$. Si, en cambio, existen tales $c_1, c_2, \ldots, c_N$ (no todos iguales a cero) que tiene lugar la igualdad (1), entonces los vectores $x_1, \ldots, x_N$ se llamarán *linealmente dependientes*. El número máximo (si existe) de vectores linealmente independientes del espacio lineal H se denomina *dimensión* del espacio citado. Un espacio que posee una infinidad de vectores linealmente independientes, se denomina de *dimensión infinita*.

El espacio H se llama *normado*, si para cada $x \in H$ viene definido un número real $\| x \|$, denominado *norma*, que satisface las siguientes condiciones:

1) $\| x \| > 0$ para $x \neq 0$; $\| x \| = 0$, si $x = 0$;
2) $\| x + y \| \leqslant \| x \| + \| y \|$ (desigualdad triangular);
3) $\| cx \| = | c | \cdot \| x \|$, donde c es un número.

Se denomina espacio *euclídeo* (*unitario*, respectivamente) el espacio lineal real de dimensión finita H (espacio lineal complejo de dimensión finita H, respectivamente), en el cual a todo par de vectores x, y se les ha puesto en correspondencia un número real (complejo) (x, y), denominado *producto escalar* de dichos vectores, con la particularidad de que se consideran cumplidas las siguientes condiciones:

Para el caso de un espacio euclídeo:

1) $(x, y) = (y, x)$ (simetría);
2) $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$ (distributividad);
3) $(\lambda x, y) = \lambda (x, y)$ (homogeneidad), donde λ es un número real cualquiera;
4) si $x \neq 0$, entonces $(x, x) > 0$.

Para el caso de un espacio unitario;

1) $(x, y) = \overline{(y, x)}$;
2) $(x_1 + x_2, y) = (x_1, y) + (x_2, y)$;
3) $(\lambda x, y) = \lambda (x, y)$ para cualquier número complejo λ;
4) si $x \neq 0$, entonces $(x, x) > 0$.

Hemos de observar que el producto escalar introducido (x, y) engendra en H la norma

$$\| x \| = \sqrt{(x, x)}. \tag{2}$$

Resulta válida aquí la desigualdad de Cauchy—Buniakovski

$$| (x, y) |^2 \leqslant (x, x) \cdot (y, y), \tag{3}$$

la cual puede escribirse, tomando en consideración (2), en la forma

$$| (x, y) | \leqslant \| x \| \cdot \| y \|.$$

2. Operadores lineales en un espacio de dimensión finita. Sea H un espacio lineal de dimensión finita provisto de producto escalar (x, y). Designemos con D cierto subespacio de H. Si a todo vector $x \in D$ se le ha puesto en correspondencia, de acuerdo con una regla determinada, el vector $y = Ax$ de H, suele decirse que en H está dado el *operador* A. El conjunto $D \subset U$ se llama *dominio de definición* del opera-

dor A y se designa $D(A)$. Un conjunto de todos los vectores del tipo $y = Ax$, $x \subset D(A)$ se denomina *campo de valores* del operador A y se denota $R(A)$. Si $D(A) = H$, dicen que el operador A está prefijado sobre H.

El operador A se llama *lineal*, si a) es aditivo, es decir, $A(x_1 + x_2) = Ax_1 + Ax_2$ para cualesquiera $x_1, x_2 \in H$; b) es homogéneo, es decir, $A(cx) = cAx$ para todo $x \in H$ y cualesquiera números c. Los requisitos a) y b) son equivalentes a la condición $A(c_1x_1 + c_2x_2) = c_1Ax_1 + c_2Ax_2$, cualesquiera que sean $x_1, x_2 \in H$ y los números c_1 y c_2.

Un operador lineal se denomina *acotado*, si existe tal constante $M > 0$ que

$$\| Ax \| \leqslant M \| x \| \qquad \text{para todo} \qquad x \in H. \tag{4}$$

La cota inferior exacta del conjunto de números M que satisfacen la condición (4) lleva el nombre de *norma* del operador A y se denota $\| A \|$. Está claro que

$$\| Ax \| \leqslant \| A \| \cdot \| x \|. \tag{5}$$

En adelante se considerarán siempre operadores lineales acotados A prefijados sobre H con el campo de valores $R(A) \subseteq H$. Tal operador A aplica H en H, lo que se escribe en la forma: A: $H \to H$.

En el espacio de dimensión finita cualquier operador lineal es acotado.

Si a cada $y \in H$ le corresponde sólo un vector $x \in H$, para el cual $Ax = y$, entonces mediante esta correspondencia queda definido un operador A^{-1}, denominado *inverso*: A^{-1}: $H \to H$. De la definición de operador inverso A^{-1} proviene que

$$A^{-1}(Ax) = x, \quad A(A^{-1}y) = y \quad \text{para cualesquiera } x, y \in H.$$

Un operador D que actúa según la regla $Dx = A(Bx)$ recibe el nombre de *producto* de los operadores A y B y se designa $D = AB$. Un operador E se denomina *operador unidad* (*idéntico*), si $Ex = x$ para todos los $x \in H$. Si existe A^{-1}, entonces $A^{-1}A = AA^{-1} = E$. Los operadores A y B se llaman *permutables* o conmutativos, si $AB = BA$.

Es evidente que A^{-1} es un operador lineal, si lo es el operador A. Resulta válida la siguiente afirmación:

Para que un operador lineal $A: H \to H$ cuente con su inverso, es necesario y suficiente que la ecuación $Ax = 0$ tenga la única solución $x = 0$.

Un operador $A^*: H \to H$ se denomina *conjugado* del operador $A: H \to H$, si

$$(Ax, y) = (x, A^*y) \quad \text{para cualesquiera} \quad x, y \in H.$$

El operador A es *autoconjugado* (*simétrico*), siempre que $A = A^*$ (o bien $(Ax, y) = (x, Ay)$ para cualesquiera x, $y \in H$). El operador lineal A se llamará: *positivo*, si $(Ax, x) > 0$ $(x \in H;\ x \neq 0)$; *definido positivo*, si $(Ax, x) \geqslant \delta \|x\|^2$ $(x \in H)$, donde $\delta > 0$ es un número; *no negativo*, si $(Ax, x) \geqslant 0$ $(x \in H)$. Cualquier operador A puede ser representado como una suma:

$$A = A_0 + A_1, \quad A_0 = \frac{1}{2}(A + A^*), \quad A_1 = \frac{1}{2}(A - A^*),$$

donde $A_0 = A_1^*$ es un operador autoconjugado y $A_1 = -A_1^*$, operador antisimétrico, para el cual en un espacio real se verifica $(A_1, x, x) = -(x, A_1x) = -(A_1x, x)$, y, por consiguiente, $(A_1x, x) = 0$. Por eso, para cualquier operador A en el espacio real H se verifica la igualdad

$$(Ax, x) = (A_0x, x) \quad \text{para todos los} \quad x \in H. \tag{6}$$

Hagamos uso de las siguientes desigualdades operacionales:

$$\begin{aligned} &A \geqslant 0, \text{ si } (Ax, x) \geqslant 0, && \text{para todos los } x \in H; \\ &A > 0, \text{ si } (Ax, x) > 0, && \text{para todos los } x \in H, x \neq 0; \\ &A \geqslant \delta E, \text{ si } (Ax, x) \geqslant \delta \|x\|^2, && \text{para todos los } x \in H, \end{aligned} \tag{7}$$

donde E es un operador *unidad*.

La desigualdad

$$B \geqslant \alpha A$$

significa que queda cumplida la condición $B - \alpha A \geqslant 0$, es decir, $((B - \alpha A)\,x, x) \geqslant 0$ (para todos los $x \in H$).

Si en un espacio real $A \neq A^*$, entonces la desigualdad $A \geqslant 0$ $(A > 0)$ será equivalente a la desigualdad $A_0 \geqslant 0$ $(A_0 > 0)$, lo que se deduce de (6)

Sea A un operador positivo. Entonces, existe un operador inverso A^{-1}: $H \to H$, siendo $A^{-1} > 0$ para $A > 0$, $(A^{-1})^* = A^{-1}$ cuando $A^* = A$. En efecto, el operador A^{-1} existe, siempre que la ecuación $Ax = 0$ tiene solamente la ecuación trivial. Admitamos que $Ax = 0$ cuando $x \neq 0$; entonces $0 = (Ax, x)$ cuando $x \neq 0$, lo que contradice la condición $A > 0$, o bien $(Ax, x) > 0$ cuando $x \neq 0$. De *este modo, si $A > 0$, entonces la ecuación $Ax = y$ tiene la* solución única.

3. Valores propios del operador lineal. Sea A un operador autoconjugado en el espacio N-dimensional H provisto de producto escalar (,). Analicemos un *problema sobre los valores propios* del operador A: se pide hallar los valores del parámetro λ (*valores propios*), para los cuales la ecuación homogénea

$$A\xi = \lambda\xi \tag{8}$$

tenga soluciones no triviales (*vectores propios*). He aquí algunas afirmaciones fundamentales del álgebra lineal sobre el problema de valores propios.

1) El operador autoconjugado A tiene N vectores propios ortonormalizados $\xi_1, \xi_2, \ldots, \xi_N$:

$$(\xi_s, \xi_m) = \delta_{sm}. \qquad \delta_{sm} = \begin{cases} 1, & s = m. \\ 0, & s \neq m. \end{cases} \tag{9}$$

2) Los valores propios correspondientes son reales y pueden disponerse en el orden de crecimiento de sus magnitudes absolutas:

$$0 \leqslant |\lambda_1| \leqslant |\lambda_2| \leqslant \ldots \leqslant |\lambda_N|. \tag{10}$$

3) Si A es un operador positivo, entonces todos los valores propios $\{\lambda_k\}$ son positivos:

$$0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N. \tag{11}$$

Efectivamente, $\lambda_s = (A\xi_s, \xi_s)/\|\xi_s\|^2 = (A\xi_s, \xi_s) > 0$, puesto que $\xi_s \neq 0$.

4) Un vector arbitrario $x \in H$ puede ser descompuesto según los vectores propios del operador $A = A^*$:

$$x = \sum_{k=1}^{N} c_k\xi_k, \quad c_k = (x, \xi_k), \tag{12}$$

quedando en este caso válida la igualdad

$$\| x \|^2 = \sum_{k=1}^{N} c_k^2. \tag{13}$$

En efecto, debido a la condición (9) de ortonormalidad del sistema $\{\xi_k\}$ tenemos

$$\| x \|^2 = (x, x) = \Big(\sum_{k=1}^{N} c_k \xi_k, \sum_{k'=1}^{N} c_{k'} \xi_{k'} \Big) =$$

$$= \sum_{k=1}^{N} \sum_{k'=1}^{N} c_k c_{k'} (\xi_k, \xi_{k'}) = \sum_{k=1}^{N} \sum_{k'=1}^{N} c_k c_{k'} \delta_{kk'} = \sum_{k=1}^{N} c_k^2.$$

5) Si $A = A^* > 0$, entonces la solución de la ecuación $Ax = f$ puede ser representada en la forma

$$x = \sum_{k=1}^{N} \frac{f_k}{\lambda_k} \xi_k, \tag{14}$$

donde $f_k = (f, \xi_k)$ es el coeficiente de Fourier de la función f. Hagamos uso de las representaciones

$$x = \sum_{k=1}^{N} c_k \xi_k, \qquad f = \sum_{k=1}^{N} f_k \xi_k$$

y escribamos

$$0 = Ax - f = \sum_{k'=1}^{N} (\lambda_{k'} c_{k'} - f_{k'}) \xi_{k'}.$$

Al multiplicar esta igualdad escalarmente por ξ_k y teniendo en cuenta que $(\xi_k, \xi_{k'}) = \delta_{kk'}$, hallemos $0 = \lambda_k c_k - f_k$, es decir, $c_k = f_k / \lambda_k$.

6) La norma de un operador autoconjugado A es igual al módulo de su valor propio máximo:

$$\| A \| = \max_{1 \leqslant k \leqslant N} |\lambda_k| = |\lambda_N|. \tag{15}$$

Efectivamente, aprovechando (12), obtenemos

$$Ax = \sum_{k=1}^{N} c_k A \xi_k = \sum_{k=1}^{N} \lambda_k c_k \xi_k,$$

y, en virtud de (10) y (13), tenemos

$$\| Ax \|^2 = \sum_{k=1}^{N} \lambda_k^2 c_k^2 \leqslant \lambda_N^2 \sum_{k=1}^{N} c_k^2 = \lambda_N^2 \| x \|^2,$$

es decir, $\| A \| \leqslant |\lambda_N|$. Esta estimación se consigue. En efecto, para $x = \xi_N$ tenemos $\| Ax \|^2 = \| A\xi_N \|^2 = \| \lambda_N \xi_N \|^2 = |\lambda_N|^2$, puesto que $\| \xi_N \|^2 = 1$. De aquí precisamente se desprende que $\| A \| = |\lambda_N|$.

7) Si $A = A^*$, entonces

$$\| A \| = \sup_{\|x\|=1} |(Ax, x)|. \tag{16}$$

8) Si $A = A^* > 0$, entonces $\lambda_1 E \leqslant A \leqslant \lambda_N E$, o bien

$$\lambda_1 \| x \|^2 \leqslant (Ax, x) \leqslant \lambda_N \| x \|^2, \quad \lambda_1 > 0, \quad x \in H. \tag{17}$$

9) Si el operador A es positivo, será definido positivo, es decir, existe una constante $\delta > 0$ tal que de la condición $A > 0$ proviene la desigualdad $A \geqslant \delta E$. Para un operador autoconjugado esta propiedad se deduce de la propiedad 8). En el caso general representemos A en forma de una suma $A = A_0 + A_1$, donde $A_0 = A_0^* > 0$, $A_1 = -A_1^*$ es un operador antisimétrico. Puesto que $(A_1 x, x) = 0$, se tiene $(Ax, x) = (A_0 x, x) > 0$. Para A_0 es cierta la propiedad 8). Suponiendo $\lambda_1 = \lambda_1 (A_0) = \delta > 0$, obtenemos $(A_0 x, x) = (Ax, x) \geqslant \delta \| x \|^2$ para todos los $x \in H$.

10) Si existe Q^{-1}, las desigualdades operacionales

$$C \geqslant 0, \qquad Q^* C Q \geqslant 0 \tag{18}$$

serán equivalentes. Esto se deduce de la identidad

$$(Q^* C Q x, x) = (CQx, Qx) = (Cy, y),$$

donde $y = Qx$, $x = Q^{-1} y$.

11) Sean A_1 y A_2 los operadores en H autoconjugados, positivos y permutables:

$$A_1 = A_1^* > 0, \qquad A_2 = A_2^* > 0, \qquad A_1 A_2 = A_2 A_1. \tag{19}$$

En este caso los operadores A_1 y A_2, la suma de ellos $A_1 + A_2$ y el producto $A_1 A_2$ tienen un sistema común de fun-

ciones propias $\{\xi_k\}$:

$$A_1\xi_k = \lambda_k^{(1)}\xi_k, \qquad A_2\xi_k = \lambda_k^{(2)}\xi_k,$$
$$\lambda(A_1 + A_2) = \lambda(A_1) + \lambda(A_2),$$
$$\lambda(A_1A_2) = \lambda(A_1)\,\lambda(A_2).$$

12) Si $A = A^* > 0$, entonces el operador $A^{-1} = (A^{-1})^* > 0$ es también autoconjugado, tiene los mismos vectores propios que el operador A, y los valores propios $\lambda(A^{-1}) = 1/\lambda(A)$.

Efectivamente, de $A\xi_k = \lambda_k\xi_k$ proviene $\xi_k = \lambda_k A^{-1}\xi_k$, es decir, $(A^{-1})\,\xi_k = (1/\lambda_k)\,\xi_k$. De aquí concluimos que las desigualdades $\lambda_1 E \leqslant A \leqslant \lambda_N E$ y $(1/\lambda_N)\,E \leqslant A^{-1} \leqslant (1/\lambda_1)\,E$ son equivalentes.

4. Problema generalizado sobre valores propios. Sea dado un operador autoconjugado positivo B. Introduzcamos un producto escalar nuevo $(x, y)_B = (Bx, y)$ y una norma $\| y \|_B = \sqrt{(By, y)}$. Un espacio H provisto de producto escalar $(x, y)_B$ recibe el nombre de *espacio energético* y se designa H_B.

Examinemos un problema generalizado sobre valores propios que consiste en buscar las soluciones no triviales v de la ecuación

$$Av = \mu Bv, \qquad v \neq 0, \tag{20}$$

donde A es un operador autoconjugado positivo.

Supongamos que los operadores A y B están representados por las matrices respectivas $A = (a_{ij})$, $B = (b_{ij})$ $(i, j = 1, 2, \ldots, N)$. La ecuación operacional (20) puede escribirse en forma de un sistema de ecuaciones algebraicas lineales

$$\sum_{j=1}^{N} a_{ij}v^{(j)} = \mu \sum_{j=1}^{N} b_{ij}v^{(j)}, \quad i = 1, 2, \ldots, N,$$

donde $v^{(1)}, \ldots, v^{(N)}$ son componentes del vector v. Para determinar los valores propios se obtiene una ecuación algebraica de N-ésimo grado

$$\det(a_{ij} - \mu b_{ij}) = 0. \tag{21}$$

Para el problema (20) son justas las propiedades análogas a las del problema corriente sobre valores propios, a saber:

existen N vectores propios ortonormalizados en el sentido del producto escalar $(x, y)_B$

$$(v_k, v_m)_B = \delta_{km}, \qquad k, m = 1, 2, \ldots, N, \tag{22}$$

a los cuales corresponden los valores propios

$$0 < \mu_1 \leqslant \ldots \leqslant \mu_N. \tag{23}$$

Por analogía con el p. 3 tenemos

$$x = \sum_{k=1}^{N} c_k v_k, \quad c_k = (x, v_k)_B,$$

$$\| x \|_B^2 = \sum_{k=1}^{N} c_k^2. \tag{24}$$

Se verifican las desigualdades operacionales

$$\mu_1 B \leqslant A \leqslant \mu_N B, \tag{25}$$

con la particularidad de que μ_N *es la norma del operador* A en H_B. Esto significa que

$$\| Ax \|_B \leqslant \| A \|_B \| x \|_B.$$

OBSERVACIÓN. Las desigualdades

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \qquad \gamma_1 > 0, \tag{26}$$

$$\gamma_1 \leqslant \mu_h \leqslant \gamma_2, \quad h = 1, 2, \ldots, N, \tag{27}$$

son equivalentes. En efecto, descompongamos un vector arbitrario $x = \sum_{k=1}^{N} c_k v_k$, hallemos $(A - \gamma B) x = \sum_{k=1}^{N} c_k (\mu_k - \gamma) B v_k$ y el producto escalar

$$((A - \gamma B) x, x) = \sum_{k=1}^{N} c_k^2 (\mu_k - \gamma)(B v_k, v_k) = \sum_{k=1}^{N} (\mu_k - \gamma) c_k^2,$$

donde γ es uno de los números γ_1 ó γ_2. Suponiendo $x = v_k$, determinemos $((A - \gamma B) v_k, v_k) = \mu_k - \gamma$. Sea $\gamma = \gamma_2$ y *supongamos cumplida la condición* $A \leqslant \gamma_2 B$; *entonces* $\mu_k \leqslant \gamma_2$. La afirmación recíproca es también cierta. Análogamente se realizan los razonamientos para $\gamma = \gamma_1$.

5. Espacios lineales de las funciones reticulares. Operadores de diferencias. En lo que sigue se examinarán sólo las funciones definidas sobre la red con nodos de números enteros:

$$\omega_N = \{i:\ i = 0, 1, \ldots, N\}.$$

Al introducir en el segmento $0 \leqslant x \leqslant 1$ los nodos $x_i = ih$, $h = 1/N$ $(i = 0, 1, \ldots, N)$, obtendremos una red uniforme de paso h como una variedad de nodos $x_i = ih$ con índices de números enteros:

$$\omega_h = \{x_i = ih:\quad i = 0, 1, \ldots, N;\quad h = 1/N\}.$$

El paso de una red a la otra es evidente y en algunos casos (bastante frecuentes) no las destinguiremos.

Denotemos con $\Omega_{N+1} = \{y_i,\ i = 0, 1, \ldots, N\}$ el espacio de funciones reticulares definidas sobre la red ω_N, con $\mathring{\Omega}_{N+1} = \{y_i,\ i = 0, 1, \ldots, N;\ y_0 = 0,\ y_N = 0\}$ el subespacio de funciones reticulares que están definidas sobre la red ω_N y se reducen a cero en los nodos de frontera de la red ω_N: $y_0 = y_N = 0$. Las funciones de $\mathring{\Omega}_{N+1}$ se designarán $\mathring{y}(i) = \mathring{y}_i$.

Veamos unos ejemplos de operadores de diferencias más simples. Para el operador de la diferencia derecha Δ tenemos

$$\Delta y_i = y_{i+1} - y_i, \qquad i = 0, 1, \ldots, N - 1;$$

aquí el dominio de definición es Ω_{N+1}, *el campo de valores* está representado por el espacio $\Omega_N^+ = \{y_i,\ i = 0, 1, \ldots, N - 1\}$ de N-ésima dimensión.

Para el operador de la diferencia izquierda ∇ tenemos

$$\nabla y_i = y_i - y_{i-1}, \qquad i = 1, 2, \ldots, N;$$

el dominio de definición es Ω_{N+1}, el campo de valores está representado por el espacio $\Omega_N^- = \{y_i,\ i = 1, 2, \ldots, N\}$.

De la fórmula

$$\Delta^2 y_{i-1} = \Delta(\Delta y_{i-1}) = \Delta(\nabla y_i) = y_{i+1} - 2y_i + y_{i-1}$$

se ve que el operador de la segunda diferencia está definido para las funciones reticulares y_i con $i = 1, 2, \ldots, N - 1$, es decir, aplica Ω_{N+1} en el espacio $\Omega_{N-1} = \{y_i,\ i = 1, 2, \ldots, N - 1\}$. La misma propiedad posee el operador de

diferencias Λ:

$$\Lambda y_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1} =$$
$$= b_i \Delta (\nabla y_i) - (b_i - a_i)(\nabla y_i) - (c_i - a_i - b_i) y_i,$$
$$i = 1, 2, \ldots, N-1,$$

es decir, $\Lambda y_i \in \Omega_{N-1}$, si $y_i \in \Omega_{N+1}$, o bien, en la notación reducida, Λ: $\Omega_{N+1} \to \Omega_{N-1}$.

Analicemos un problema de contorno en diferencias

$$\Lambda y_i = -f_i, \qquad i = 1, 2, \ldots, N-1,$$
$$y_0 = \mu_1, \qquad y_N = \mu_2 \tag{28}$$

y escribámosla en la forma matricial:

$$AY = \Phi, \tag{29}$$

donde $\Phi = (f_1 + a_1\mu_1, f_2, \ldots, f_{N-2}, f_{N-1} + b_{N-1}\mu_2)$ es el vector conocido e $Y = (y_1, y_2, \ldots, y_{N-2}, y_{N-1})$ es un vector desconocido, ambos de dimensión $N-1$; A es una matriz tridiagonal cuadrada de dimensión $(N-1) \times \times (N-1)$:

$$A = -\begin{bmatrix} -c_1 & b_1 & \ldots & 0 \\ a_2 & -c_2 & b_2 \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 \ldots 0 & & a_{N-1} & -c_{N-1} \end{bmatrix}.$$

Al comparar (28) y (29), vemos que se puede escribir

$$\tilde{\Lambda} y_i = -\varphi_i, \; i = 1, 2, \ldots, N-1,$$
$$\tilde{\Lambda} y_1 = -c_1 y_1 + b_1 y_2, \quad \varphi_1 = f_1 + a_1\mu_1,$$
$$\tilde{\Lambda} y_i = \Lambda y_i, \; \varphi_i = f_i, \; i = 2, 3, \ldots, N-2,$$
$$\tilde{\Lambda} y_{N-1} = a_{N-1} y_{N-2} - c_{N-1} y_{N-1}, \; \varphi_{N-1} = f_{N-1} + b_{N-1}\mu_2. \tag{28'}$$

El operador de diferencias $\tilde{\Lambda}$ aplica Ω_{N-1} en Ω_{N-1}. No es difícil observar que $\Lambda \tilde{y}_i = \tilde{\Lambda} \dot{y}_i$. En lugar de (28′) obtendremos

$$\Lambda \dot{y}_i = -\varphi_i, \qquad i = 1, 2, \ldots, N-1.$$

Introduzcamos ahora el operador A correspondiente a la matriz (29), suponiendo

$$Ay_i = -\tilde{\Lambda} y_i = -\mathring{\Lambda} y_i, \qquad i = 1, 2, \ldots, N-1.$$

En tal caso, en vez del problema de contorno en diferencias (28) obtendremos una ecuación operacional

$$Ay = \varphi,$$

donde $A: \Omega_{N-1} \to \Omega_{N-1}$, $\varphi \in \Omega_{N-1}$, es decir, el operador A actúa de Ω_{N-1} en Ω_{N-1}. Es evidente que A será un operador lineal. *Ha de notarse que también puede considerarse* (teniendo en cuenta que $Ay = -\Lambda \mathring{y}$), que A aplica $\mathring{\Omega}_{N-1}$ en Ω_{N-1}.

En el espacio $H = \Omega_{N-1}$ se puede introducir un producto escalar

$$(y, v) = \frac{1}{N} \sum_{i=1}^{N-1} y_i v_i$$

y una norma

$$\|y\| = \sqrt{(y, y)}.$$

Si se estudian el segundo ($\varkappa_1 = \varkappa_2 = 1$) o el tercero ($\varkappa_1 \neq 0$, $\varkappa_2 \neq 0$) de los problemas de contorno (véase (1) del § 3), la matriz A será cuadrada de dimensión $(N+1) \times \times (N+1)$ y el operador A se definirá de la manera siguiente:

$$Ay_i = -\Lambda y_i = -(b_i y_{i+1} - c_i y_i + a_i y_{i-1}),$$
$$i = 1, 2, \ldots, N-1,$$
$$Ay_0 = -(\varkappa_1 y_1 - y_0), \quad Ay_N = -(y_N - \varkappa_2 y_{N-1}).$$

En este caso el operador A aplica el espacio de funciones reticulares $H = \Omega_{N+1}$ en sí mismo: $A: H \to H$.

En adelante se analizará el primer problema de contorno para una ecuación en diferencias de segundo orden; en este caso, según lo mostrado más arriba, $H = \Omega_{N-1}$.

6. Fórmulas de Green de diferencias. Examinemos un operador de diferencias L:

$$Ly_i = b_i y_{i+1} - c_i y_i + a_i y_{i-1}, \quad i = 1, \ldots, N-1. \qquad (30)$$

Si $b_i \neq a_{i+1}$, la matriz correspondiente no será simétrica. Es simétrica sólo en el caso

$$b_i = a_{i+1}, \qquad i = 1, 2, \ldots, N-1. \tag{31}$$

Al tomar en consideración esta condición, escribimos Ly_i en la forma siguiente:

$$\begin{aligned} Ly_i &= a_{i+1}y_{i+1} - c_iy_i + a_iy_{i-1} = \\ &= a_{i+1}(y_{i+1} - y_i) - a_i(y_i - y_{i-1}) - (c_i - a_i - \\ &- a_{i+1})\, y_i = a_{i+1}\nabla y_{i+1} - a_i \nabla y_i - (c_i - a_i - \\ &- a_{i+1})\, y_i = \Delta(a_i \nabla y_i) - (c_i - a_i - a_{i+1})\, y_i. \end{aligned} \tag{32}$$

Dividamos el segmento $[0, 1]$ con los puntos x_i en N partes iguales, hagamos $y(x_i) = y_i = y(i)$ e introduzcamos las designaciones que siempre se usarán en lo sucesivo:

$$\begin{gathered} h = \frac{1}{N}, \quad x_i = ih, \quad i = 0, 1, \ldots, N, \ x_0 = 0, \quad x_N = 1, \\ y_{x,i} = \frac{\Delta y_i}{h} = \frac{y_{i+1} - y_i}{h}, \quad y_{\bar{x},i} = \frac{\nabla y_i}{h} = \frac{y_i - y_{i-1}}{h}, \\ y_{\bar{x}x,i} = y_{\bar{x}x}(i) = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = \frac{\Delta(\nabla y_i)}{h^2}. \end{gathered} \tag{33}$$

Dividamos la expresión (32) por h^2 y obtendremos un operador de diferencias

$$\begin{gathered} \Lambda y_i = (ay_{\bar{x}})_{x,i} - d_iy_i, \\ d_i = \frac{1}{h^2}(c_i - a_i - a_{i+1}), \quad i = 1, \ldots, N-1. \end{gathered} \tag{34}$$

En el § 1 se ha obtenido la fórmula de sumación por partes

$$\sum_{i=0}^{N-1} y_i \Delta v_i = -\sum_{i=1}^{N} v_i \nabla y_i + (yv)_N - (yv)_0. \tag{35}$$

Haciendo uso de las designaciones (33), escribamos esta fórmula en la forma

$$\sum_{i=0}^{N-1} y_i v_{x,i}\, h = -\sum_{i=1}^{N} v_i y_{\bar{x},i}\, h + (yv)_N - (yv)_0, \tag{36}$$

puesto que $\sum_{i=0}^{N-1} y_i \Delta v_i = \sum_{i=0}^{N-1} y_i \left(\frac{\Delta v_i}{h}\right) h = \sum_{i=0}^{N-1} y_i v_{x,i}\, h.$

Para que la exposición ulterior sea más cómoda, introduzcamos en el primer miembro de (36) la sumación entre $i = 1$ e $i = N - 1$; esto nos conduce a la fórmula

$$\sum_{i=1}^{N-1} y_i v_{x,i} h = -\sum_{i=1}^{N} v_i y_{\bar{x},i} h + (yv)_N - y_0 v_1. \qquad (37)$$

Sustituyamos aquí $v_i = a_i z_{\bar{x},i}$; se obtendrá

$$\sum_{i=1}^{N-1} y_i (a z_{\bar{x}})_{x,i} h = -\sum_{i=1}^{N} a_i y_{\bar{x},i} z_{\bar{x},i} h + \\ + (a y z_{\bar{x}})_N - y_0 (a z_{\bar{x}})_1. \qquad (38)$$

Esta es la *primera fórmula de Green de diferencias*. Cambiemos de lugar en ella y_i y z_i:

$$\sum_{i=1}^{N-1} z_i (a y_{\bar{x}})_{x,i} h = -\sum_{i=1}^{N} a_i z_{\bar{x},i} y_{\bar{x},i} h + \\ + (a y_{\bar{x}} z)_N - z_0 (a y_{\bar{x}})_1. \qquad (38')$$

Al sustraer (38') de (38) obtenemos la *segunda fórmula de Green de diferencias*

$$\sum_{i=1}^{N-1} y_i (a z_{\bar{x}})_{x,i} h = \sum_{i=1}^{N-1} z_i (a y_{\bar{x}})_{x,i} h + a_N (y z_{\bar{x}} - z y_{\bar{x}})_N - \\ - (y_0 (a z_{\bar{x}})_1 - z_0 (a y_{\bar{x}})_1). \qquad (39)$$

Si están cumplidas las condiciones

$$y_0 = z_0 = 0, \qquad y_N = z_N = 0, \qquad (40)$$

es decir, si $y = \overset{\circ}{y}$, $z = \overset{\circ}{z} \in \overset{\circ}{\Omega}_{N+1}$, entonces en el segundo miembro de la igualdad (39) dos últimos sumandos se anulan y

$$\sum_{i=1}^{N-1} \overset{\circ}{y}_i (a \overset{\circ}{z}_{\bar{x}})_{x,i} h = \sum_{i=1}^{N-1} \overset{\circ}{z}_i (a \overset{\circ}{y}_{\bar{x}})_{x,i} h. \qquad (41)$$

Al sustraer de ambos miembros de la identidad (41) la suma $\sum_{i=1}^{N-1} d_i \overset{\circ}{y}_i \overset{\circ}{z}_i h$, obtenemos la *segunda fórmula de Green* para

$y, z \in \dot{\Omega}_{N+1}$:

$$\sum_{i=1}^{N-1} \dot{y}_i \Lambda \dot{z}_i h = \sum_{i=1}^{N-1} \dot{z}_i \Lambda \dot{y}_i h \qquad (42)$$

para el operador de diferencias

$$\Lambda \dot{y}_i = (a\dot{y}_{\bar{x}})_{x,i} - d_i \dot{y}_i, \text{ cualquiera que sea } \dot{y} \in \dot{\Omega}_{N+1}. \qquad (43)$$

Sea $H = \Omega_{N-1}$ un espacio de funciones reticulares y_i, prefijadas para $i = 1, 2, \ldots, N-1$, con el producto escalar

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h$$

y la norma

$$\| y \| = \sqrt{(y, y)}.$$

Introduzcamos el operador A:

$$Ay = -\Lambda \dot{y}, \qquad y \in H. \qquad (44)$$

Entonces la segunda fórmula de Green puede anotarse en la forma

$$(y, Az) = (Ay, z). \qquad (45)$$

Esta fórmula expresa la propiedad de autoconjugación del operador A: $A^* = A$ y, por lo tanto, $\Lambda^* = \Lambda$. Cuando $\dot{z} = \dot{y} \in \dot{\Omega}_{N+1}$, la primera fórmula de Green (38) nos da:

$$-\sum_{i=1}^{N-1} \dot{y}_i (a\dot{y}_{\bar{x}})_{x,i} h = \sum_{i=1}^{N} a_i (\dot{y}_{\bar{x},i})^2 h > 0$$

$$\text{para } \dot{y}_i \not\equiv 0, \ a_i > 0, \qquad (46)$$

(ya que $\dot{y}_0 = \dot{y}_N = 0$, (46) puede ser igual a cero sólo en el caso en que $\dot{y}_i \equiv 0$ $(i = 1, \ldots, N-1)$). Teniendo presente la definición del operador A, hallamos

$$(Ay, y) = \sum_{i=1}^{N} a_i (y_{\bar{x},i})^2 h + \sum_{i=1}^{N} d_i y_i^2 h > 0, \quad a_i > 0,$$

$$d_i \geqslant 0. \qquad (47)$$

De este modo, el operador de diferencias A, definido por las fórmulas (43), (44), es autoconjugado y positivo: $A = A^* > 0$, siempre que

$$a_i > 0, \quad d_i \geqslant 0, \quad i = 1, 2, \ldots, N-1, \quad a_N > 0. \tag{48}$$

7. Condición de autoconjugación del operador de diferencias de segundo orden. Nos hemos convencido de que la condición (31) es suficiente para que el operador de diferencias (30) sea autoconjugado en el espacio $H = \mathring{\Omega}_{N+1}$. Mostremos que la condición (31) es necesaria para que sea autoconjugado L. Representemos L en forma de una suma:

$$Ly_i = L_1 y_i + L_2 y_i,$$

$$L_1 y_i = a_{i+1}(y_{i+1} - y_i) - a_i(y_i - y_{i-1}) - (c_i - a_i - b_i)y_i,$$

$$L_2 y_i = (b_i - a_{i+1})\, y_{i+1}.$$

Como ya se ha mostrado en el punto antecedente, el operador $L_1 y_i = h^2 \Lambda y_i$, $\Lambda y_i = (a y_{\bar{x}})_{x,i} - d_i y_i$, es autoconjugado en el espacio $H = \mathring{\Omega}_{N+1}$ ó en $H = \Omega_{N-1}$ con el producto escalar $(y, v) = \sum_{i=1}^{N-1} y_i v_i h$. Por eso podemos escribir:

$$\left(\frac{1}{h^2} L\mathring{y}, \mathring{v}\right) - \left(\mathring{y}, \frac{1}{h^2} L\mathring{v}\right) =$$

$$= (\Lambda\mathring{y}, \mathring{v}) - (\mathring{y}, \Lambda\mathring{v}) + \left(\frac{1}{h^2} L_2\mathring{y}, \mathring{v}\right) - \left(\mathring{y}, \frac{1}{h^2} L_2\mathring{v}\right) =$$

$$= \sum_{i=1}^{N} \frac{1}{h^2}(b_i - a_{i+1})(y_{i+1}v_i - y_i v_{i+1})\, h.$$

De aquí se ve que $(L\mathring{y}, \mathring{v}) = (\mathring{y}, L\mathring{v})$, es decir, $L = L^*$ sólo a la condición de que

$$\sum_{i=1}^{N-1} (b_i - a_{i+1})(y_{i+1}v_i - y_i v_{i+1})\, h = 0. \tag{49}$$

Por ser arbitrarias y_i y v_i, podemos tomar $y_i = \delta_{i, i_0+1}$, $v_i = \delta_{i, i_0}$, donde i_0 es un nodo fijo cualquiera ($i_0 = 1, 2, \ldots, N-1$), mientras que δ_{i, i_0} es el símbolo de Kronecker.

Obtenemos, pues, $y_{i+1}v_i - y_i v_{i+1} = \delta_{i,i_0}$, y la condición (49) nos da $b_{i_0} = a_{i_0+1}$. Con esto queda demostrada la necesidad de la condición (31).

Se debe notar que la ecuación

$$Ly_i = -f_i \tag{50}$$

puede ser reducida a la forma

$$\tilde{L}y_i = \Delta (A_i \nabla y_i) - D_i y_i = -F_i, \tag{51}$$

donde $\tilde{L}$ es un operador autoconjugado. En efecto, multipliquemos ambos miembros de la ecuación (50) por $\mu_i \neq 0$:

$$\tilde{L}y_i = \mu_i a_i y_{i-1} - \mu_i c_i y_i + b_i \mu_i y_{i+1} = -\mu_i f_i$$

y exijamos que para la ecuación obtenida se cumpla la condición (31), es decir,

$$b_i \mu_i = (\mu a)_{i+1} = a_{i+1}\mu_{i+1} = A_{i+1}.$$

De aquí obtenemos $\mu_{i+1} = \left(\frac{b_i}{a_{i+1}}\right)\mu_i = \mu_1 \prod_{k=1}^{i} b_k/a_{k+1}$ y la ecuación (51), donde $A_i = a_i\mu_i$, $D_i = \mu_i(c_i - a_i - b_i)$, $F_i = \mu_i f_i$.

8. Valores propios del operador de diferencias de segundo orden. Examinemos un problema de diferencias sobre valores propios:

$$(ay_{\bar{x}})_{x,i} - d_i y_i + \lambda y_i = 0, \quad i = 1, 2, \ldots, N-1$$
$$y_0 = y_N = 0, \tag{52}$$

o bien $Ay = \lambda y$, $y \in \Omega_{N-1}$, donde A se determina por la igualdad (44). El operador A es autoconjugado y positivo, razón por la cual a él se refiere todo lo dicho en el p. 4.

En el caso más simple, $a_i = 1$, $d_i = 0$, los valores propios y los vectores propios pueden hallarse en la forma explícita. Así pues, se requiere encontrar soluciones no triviales de la ecuación homogénea con condiciones de contorno homogéneas

$$y_{\bar{x}x,i} + \lambda y_i = 0, \quad i = 1, 2, \ldots, N-1, \quad hN = 1,$$
$$y_0 = 0, \quad y_N = 0, \quad y_i \not\equiv 0. \tag{53}$$

Escribamos la ecuación (53) en la forma siguiente

$$y_{i-1} - 2\cos\alpha y_i + y_{i-1} = 0, \quad 2\cos\alpha = 2 - \lambda h^2. \tag{54}$$

La solución general de esta ecuación tiene por expresión

$$y_i = c_1 \cos i\alpha + c_2 \operatorname{sen} i\alpha. \tag{55}$$

Exigimos que se cumplan las condiciones de contorno: $y_0 = c_1 = 0$, $y_N = c_2 \operatorname{sen} N\alpha = 0$. Como se busca una solución no trivial, entonces $c_2 \neq 0$ y $\operatorname{sen} N\alpha = 0$, es decir, $N\alpha = m\pi$ $(m = 0, 1, 2, \ldots)$, $\alpha = \alpha_m = m\pi/N = m\pi h$. De la relación $2\cos\alpha = 2 - \lambda h^2$ encontramos

$$\lambda h^2 = 2(1 - \cos\alpha) = 4\operatorname{sen}^2\frac{\alpha}{2},$$

$$\lambda = \lambda_m = \frac{4}{h^2}\operatorname{sen}^2\frac{\pi m h}{2}. \tag{56}$$

A este valor de λ_m le corresponde una función propia

$$y_m(i) = c \operatorname{sen} \pi m x_i, \; c \neq 0, \; x_i = ih, \; i = 0, 1, 2, \ldots, N \tag{57}$$

definida con una exactitud de hasta un factor constante arbitrario. No es difícil notar que

$$y_N(i) = c \operatorname{sen} \pi N x_i = c \operatorname{sen} \pi i = 0, \; i = 0, 1, 2, \ldots,$$

$$y_{N+1}(i) = c \operatorname{sen} \pi (N+1) x_i = c \operatorname{sen} [\pi N x_i + \pi x_i] =$$

$$= c \operatorname{sen} \pi x_i \cos \pi i = (-1)^i y_1(i),$$

$$y_{N+m+1}(i) = (-1)^i y_m(i), \quad m = 1, 2, \ldots, N-1.$$

Por consiguiente, para $m < N$, sólo las funciones $y_m(i)$ son linealmente independientes. Así pues, se ha encontrado la solución no trivial (funciones propias $y_m(i)$ que corresponden a los valores propios λ_m).

Elijamos el factor c de un modo tal que la norma de las funciones $y_m(i)$ sea igual a la unidad: $\| y_m(i) \| = c \, \| \operatorname{sen} \pi m x_i \| = 1$, $c > 0$. Con este fin se debe calcular

$$\| \operatorname{sen} \pi m x_k \|^2 = \sum_{k=1}^{N-1} h \operatorname{sen}^2 \pi m x_k = \frac{1}{2} \sum_{k=1}^{N-1} h (1 - \cos 2\pi m x_k).$$

Al denotar $\alpha = 2\pi mh$ y sustituir $\cos 2\pi m x_k = \cos \alpha k = \operatorname{Re} e^{i\alpha k}$, llegamos a que

$$\sum_{k=1}^{N-1} h \cos 2\pi m x_k = \operatorname{Re} \sum_{k=1}^{N-1} h e^{i\alpha k} = h \operatorname{Re} \frac{e^{i\alpha} - e^{i\alpha N}}{1 - e^{i\alpha}} = -h,$$

$$\| \operatorname{sen} \pi m x_k \|^2 = \frac{(N-1)h}{2} - \frac{1}{2} \sum_{k=1}^{N-1} h \cos 2\pi m x_k = \frac{Nh}{2} = \frac{1}{2},$$

$$\| \operatorname{sen} \pi m x \| = 1/\sqrt{2};$$

por consiguiente, $c = \sqrt{2}$. De este modo, la función

$$y_m(i) = \sqrt{2} \operatorname{sen} \pi m x_i \tag{58}$$

está normalizada hacia la unidad.

Las funciones propias $y_s(i)$ e $y_m(i)$, correspondientes a los diferentes valores propios λ_s y λ_m, son ortogonales en el sentido del producto escalar

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

El problema (53) constituye un caso particular del problema (8) con el operador $Ay(i) = -y_{\bar{x}\overset{\circ}{x}}(i)$. Dicho operador es, evidentemente, autoconjugado y positivo, puesto que

$$(Ay, y) = \sum_{i=1}^{N-1} (y_{\bar{x},i})^2 h > 0.$$

Por esta razón todo lo dicho en el p. 3 queda vigente también en el caso dado.

Los valores propios λ_s crecen a medida que crece s, puesto que $\operatorname{sen} \frac{\pi h}{2} s < \operatorname{sen} \frac{\pi h}{2}(s+1) < 1$ para $s \leqslant N$. El valor propio mínimo es $\lambda_1 = \frac{4}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}$. El valor propio máximo es igual a $\lambda_{N-1} = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$, ya que $\operatorname{sen} \frac{\pi h}{2} (N-1) = \operatorname{sen} \left(\frac{\pi}{2} - \frac{\pi h}{2} \right) = \cos \frac{\pi h}{2}$.

Escribiendo λ_1 en la forma $\lambda_1 = \pi^2 \left(\frac{\operatorname{sen}\xi}{\xi}\right)^2$, $\xi = \pi h/2 = \pi h/2 \leqslant \pi/4$, y teniendo presente que sen ξ/ξ decrece y tiene mínimo para $\xi = \pi/4$, obtenemos $\lambda_1 \geqslant 8$ para $h \leqslant 1/2$.

Para λ_{N-1} tenemos una estimación $\lambda_{N-1} < 4/h^2$, y, por consiguiente,

$$8 < \lambda_k < 4/h^2, \qquad k = 1, 2, \ldots, N-1.$$

§ 5. Principio del máximo para las ecuaciones en diferencias

1. Principio del máximo y sus corolarios. Para las ecuaciones en diferencias de segundo orden con coeficientes positivos

$$Ly_i = a_i y_{i-1} - c_i y_i + b_i y_{i+1} = -\varphi_i,$$
$$i = 1, 2, \ldots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2, \tag{1}$$
$$a_i > 0, \quad b_i > 0, \quad c_i \geqslant a_i + b_i, \quad i = 1, 2, \ldots, N-1 \tag{2}$$

tiene lugar el siguiente principio del máximo.

TEOREMA 1 (*principio del máximo*). *Supongamos que un operador de diferencias L está definido por las fórmulas* (1), (2). *Si para una función y_i, prefijada sobre la red $\bar{\omega}$ y diferente de una constante* ($1 \leqslant i \leqslant N-1$), *se cumple la condición* $Ly_i \geqslant 0$ ($Ly_i \leqslant 0$) para todo $i = 1, 2, \ldots, N-1$, *entonces dicha función no puede tomar el valor positivo máximo (negativo mínimo) en los nodos interiores de la red.*

DEMOSTRACIÓN. Sea $Ly_i \geqslant 0$ $(i = 1, 2, \ldots, N-1)$. Supongamos que el teorema no es cierto e y_i alcanza su valor positivo máximo en un nodo interior $i = i_*$, $1 \leqslant i_* \leqslant N-1$: $y_{i_*} = \max\limits_{0 \leqslant i \leqslant N} y_i = M_0 > 0$. Como $y_i \not\equiv$ const, se encontrará un nodo interior i_0 (i_0 puede coincidir con i_*) en el cual $y_{i_0} = y_{i_*} = M_0 > 0$, y en uno de los nodos vecinos, por ejemplo, en el nodo $i = i_0 - 1$, se verifica la desigualdad rigurosa $y_{i_0-1} < y_{i_0}$. Escribamos la expresión para Ly_i en la forma $Ly_i = b_i(y_{i+1} - y_i) - a_i(y_i - y_{i-1}) - (c_i - a_i - b_i)y_i$. En el nodo $i = i_0$

tenemos

$$Ly_{i_0} = b_{i_0}(y_{i_0+1} - y_{i_0}) - a_{i_0}(y_{i_0} - y_{i_0-1}) - \\ - (c_{i_0} - a_{i_0} - b_{i_0})\, y_{i_0} < 0,$$

lo que contradice la suposición $Ly_i \geqslant 0$ para cualquiera de los $i = 1, 2, \ldots, N-1$, incluso para $i = i_0$. La primera afirmación del teorema queda demostrada. La segunda afirmación se demuestra análogamente (basta sustituir y_i por $-y_i$ y aprovechar la afirmación que acabamos de demostrar).

COROLARIO 1. *Si se cumplen las condiciones* (2), es decir, *si* $Ly_i \leqslant 0$ $(i = 1, 2, \ldots, N-1)$, $y_0 \geqslant 0$, $y_N \geqslant 0$, *entonces* $y_i \geqslant 0$ $(i = 0, 1, \ldots, N)$.

Si $Ly_i \geqslant 0$, $y_0 \leqslant 0$, $y_N \leqslant 0$, *entonces* $y_i \leqslant 0$ $(i = 0, 1, \ldots, N)$.

DEMOSTRACION. Supongamos que $Ly_i \leqslant 0$ e $y_i < 0$ por lo menos en uno de los nodos interiores $i = i_*$; entonces y_i alcanza el valor negativo mínimo en el nodo interior, lo que es imposible en virtud del principio del máximo.

COROLARIO 2. *Si* $\varphi_i \geqslant 0$, $\mu_1 \geqslant 0$, $\mu_2 \geqslant 0$, *entonces la solución del problema* (1)—(2) *es no negativa:* $y_i \geqslant 0$ $(i = 0, 1, \ldots, N)$.

COROLARIO 3. *Si quedan cumplidas las condiciones* (2), *el problema*

$$Ly_i = 0, \quad i = 1, 2, \ldots, N-1, \quad y_0 = 0, \quad y_N = 0 \qquad (3)$$

tiene sólo una solución trivial y el problema (1), (2) *es resoluble unívocamente, cualesquiera que sean* φ_i, μ_1, μ_2.

DEMOSTRACION. Suponiendo que la solución y_i del problema (3) es diferente de cero por lo menos en un solo punto $i = i_*$, llegamos a una contradicción con el principio del máximo: si $y_{i_*} > 0$ $(y_{i_*} < 0)$, entonces y_i alcanza el valor máximo positivo (mínimo negativo) en cierto punto interior $i = i_0$, lo que es imposible. Por consiguiente, $y_i \equiv 0$.

TEOREMA 2. (*teorema de comparación*). *Supongamos que* y_i *es la solución del problema* (1), (2) *e* $\bar{y}_i$, *la solución del problema*

$$L\bar{y}_i = -\bar{\varphi}_i, \quad i = 1, 2, \ldots, N-1, \quad \bar{y}_0 = \bar{\mu}_1, \\ \bar{y}_N = \mu_2$$

y, además, admitamos cumplidas las condiciones

$$|\varphi_i| \leqslant \bar{\varphi}_i, \quad |\mu_1| \leqslant \bar{\mu}_1, \quad |\mu_2| \leqslant \bar{\mu}_2.$$

En este caso resulta válida la estimación

$$|y_i| \leqslant \bar{y}_i \text{ para todo } i = 0, 1, \ldots, N.$$

DEMOSTRACIÓN. En virtud del corolario 2, tenemos $\bar{y}_i \geqslant 0$. Para la diferencia $\bar{y}_i - y_i$ y para la suma $\bar{y}_i + y_i$ obtenemos una ecuación del tipo (1) con los segundos miembros $\bar{\varphi}_i - \varphi_i \geqslant 0$, $\bar{\mu}_1 - \mu_1 \geqslant 0$, $\bar{\mu}_2 - \mu_2 \geqslant 0$ y $\bar{\varphi}_i + \varphi_i \geqslant 0$, $\bar{\mu}_1 + \mu_1 \geqslant 0$, $\bar{\mu}_2 + \mu_2 \geqslant 0$, respectivamente. Puesto que $\bar{\varphi}_i \pm \varphi_i \geqslant 0$ y $\bar{\mu}_\alpha \pm \mu_\alpha \geqslant 0$ $(\alpha = 1, 2)$, entonces, debido al corolario 2, $\bar{y}_i - y_i \geqslant 0$, $\bar{y}_i + y_i \geqslant 0$, de lo que se deduce que $-\bar{y}_i \leqslant y_i \leqslant \bar{y}_i$, $|y_i| \leqslant \bar{y}_i$, lo que se trataba de demostrar.

La función $\bar{y}_i$ se denomina *mayorante* para la solución del problema (1), (2).

2. Estimación de la solución del problema de contorno. Representemos la solución del problema de contorno (1), (2) en forma de una suma $y_i = y_i^{(1)} + y_i^{(2)}$, donde $y_i^{(1)}$ es la solución de la ecuación no homogénea con condiciones de contorno homogéneas:

$$Ly_i = -\varphi_i, \quad i = 1, 2, \ldots, N-1, \quad y_0 = y_N = 0, \tag{4}$$

mientras que $y_i^{(2)}$ representa la solución de la ecuación homogénea con condiciones de contorno no homogéneas

$$Ly_i = 0, \quad i = 1, 2, \ldots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2. \tag{5}$$

Demostremos que para $y_i^{(2)}$ es justa la estimación

$$\max_{0 \leqslant i \leqslant N} |y_i^{(2)}| \leqslant \max(|\mu_1|, |\mu_2|). \tag{6}$$

Sea $\bar{y}_i$ la solución del problema

$$L\bar{y}_i = 0, \quad i = 1, 2, \ldots, N-1,$$

$$\bar{y}_0 = \bar{y}_N = \bar{\mu}, \quad \bar{\mu} = \max(|\mu_1|, |\mu_2|).$$

Entonces, de acuerdo con el teorema de comparación, $|y_i^{(2)}| \leqslant |\overline{y}_i|$, mientras que, en virtud del principio del máximo $\max\limits_{0 \leqslant i \leqslant N} |\overline{y}_i| \leqslant \mu$, puesto que $\overline{y}_i \geqslant 0$ puede alcanzar el valor positivo máximo sólo en la frontera, es decir, cuando $i = 0$ o $i = N$.

No es difícil demostrar que la magnitud $\max\limits_{0 \leqslant i \leqslant N} |y_i|$ es una norma. La norma suele designarse con el símbolo $\| y \|_c$. De este modo, hemos obtenido la estimación $\| y^{(2)} \|_C \leqslant \leqslant \max(|\mu_1|, |\mu_2|)$.

TEOREMA 3. *Supongamos que están cumplidas las condiciones*

$$|a_i| > 0, \quad |b_i| > 0, \quad \overline{d}_i = |c_i| - |a_i| - |b_i| > 0$$
$$i = 1, 2, \ldots, N-1. \tag{7}$$

Entonces, para la solución del problema (4) *es justa la estimación*

$$\| y \|_c \leqslant \| \varphi/\overline{d} \|_c. \tag{8}$$

DEMOSTRACION. Con el fin de demostrar la citada afirmación escribamos (4) en la forma

$$c_i y_i = a_i y_{i-1} + b_i y_{i+1} + \varphi_i. \tag{4'}$$

Supongamos que $|y_i|$ alcanza su valor máximo $|y_{i_0}| > > 0$ cuando $i = i_0$ $(0 < i_0 < N)$, de suerte que $|y_i| \leqslant \leqslant |y_{i_0}|$ para cualquier $i = 0, 1, \ldots, N$. Entonces, de (4') se deduce para $i = i_0$:

$$|c_{i_0}|\,|y_{i_0}| = |a_{i_0} y_{i_0-1} + b_{i_0} y_{i_0+1} + \varphi_{i_0}| \leqslant |a_{i_0}|\,|y_{i_0-1}| +$$
$$+ |b_{i_0}|\,|y_{i_0+1}| + |\varphi_{i_0}| \leqslant$$
$$\leqslant (|a_{i_0}| + |b_{i_0}|)\,|y_{i_0}| + |\varphi_{i_0}|,$$

$$(|c_{i_0}| - |a_{i_0}| - |b_{i_0}|)\,|y_{i_0}| \leqslant |\varphi_{i_0}|, \quad |y_{i_0}| \leqslant \frac{|\varphi_{i_0}|}{\overline{d}_{i_0}} \leqslant \left\| \frac{\varphi}{\overline{d}} \right\|_c.$$

Con esto queda demostrada la estimación (8).

OBSERVACION. Si la condición $d_i = c_i - a_i - b_i > 0$ o $\overline{d}_i = |c_i| - |a_i| - |b_i| > 0$ no se cumple, por ejemplo,

$$d_i = c_i - a_i - b_i \geqslant 0, \; a_i > 0, \; b_i > 0, \; i = 1, 2, \ldots, N-1, \tag{9}$$

es decir, d_i puede reducirse a cero en ciertos nodos, entonces el teorema 3 no puede ser aplicado. En este caso, con el fin de estimar la solución y_i del problema (4), se puede proceder de la manera siguiente. Representemos y_i en forma de una suma $y_i = v_i + w_i$, donde w_i es la solución del problema

$$\overset{\bullet}{L}w_i = b_i (w_{i+1} - w_i) - a_i (w_i - w_{i-1}) = -\varphi_i,$$
$$i = 1, 2, \ldots, N-1, \quad w_0 = w_N = 0. \qquad (10)$$

Entonces, v_i se determina partiendo de las condiciones

$$Lv_i = b_i (v_{i+1} - v_i) - a_i (v_i - v_{i-1}) - d_i v_i = -d_i w_i,$$
$$i = 1, 2, \ldots, N-1, \quad v_0 = v_N = 0. \qquad (11)$$

Se puede convencerse de esto sumando término a término las ecuaciones (10) y (11). La función w_i puede estimarse inmediatamente (véase el cap. IV, § 3), al escribirla en la forma explícita, mientras que para la estimación de v_i necesitaremos el

TEOREMA 4. *Para resolver el problema* (11) *bajo las condiciones* (9) *es válida la estimación*

$$\| v \|_c \leqslant \| w \|_c. \qquad (12)$$

DEMOSTRACIÓN. Si $d_i \equiv 0$, entonces, en virtud del corolario 3, $v_i \equiv 0$, y la estimación (12) queda cumplida. Sea $d_i \not\equiv 0$ siquiera en un solo punto. Construyamos la mayorante $\bar{v}_i$ como una solución del problema

$$L\bar{v}_i = -d_i \mid w_i \mid, \quad i=1, 2, \ldots, N-1, \quad \bar{v}_0 = \bar{v}_N = 0.$$

Supongamos que $\bar{v}_i \geqslant 0$ alcanza su valor máximo para $i = i_0$; entonces $\bar{v}_{i_0+1} - \bar{v}_{i_0} \leqslant 0$, $\bar{v}_{i_0} - \bar{v}_{i_0-1} \geqslant 0$, y de (4) proviene

$$d_{i_0}\bar{v}_{i_0} \leqslant -b_{i_0} (\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0} (\bar{v}_{i_0} - \bar{v}_{i_0-1}) + d_{i_0}\bar{v}_{i_0} =$$
$$= d_{i_0} \mid w_{i_0} \mid.$$

Si $d_{i_0} > 0$, entonces $\bar{v}_{i_0} < \mid w_{i_0} \mid$ y obtenemos en seguida la estimación (12), puesto que $\mid v_i \mid \leqslant \bar{v}_i$. Si $d_{i_0} = 0$, la ecuación (11) tomará la forma $-b_{i_0} (\bar{v}_{i_0+1} - \bar{v}_{i_0}) + a_{i_0} (\bar{v}_{i_0} - \bar{v}_{i_0-1}) = 0$, y de esta última se deduce que $\bar{v}_{i_0+1} = \bar{v}_{i_0-1} = \bar{v}_{i_0}$. Por

cuanto $\bar{v}_i \not\equiv$ const, existe tal punto $i = i_1$ en el cual $\bar{v}_{i_1} = \bar{v}_{i_0}$, y en el punto vecino, por ejemplo, $i = i_1 + 1$, $\bar{v}_{i_1+1} < \bar{v}_{i_1}$; entonces aquí $d_{i_1} \neq 0$ y obtenemos, pues, el caso analizado más arriba: $\bar{v}_{i_1} = \|\bar{v}\|_c \leqslant |w_{i_0}| \leqslant \|w\|_c$.

3. Estimación de la solución de una ecuación en diferencias con ayuda de las fórmulas de factorización. Para el caso en que $b_i = a_{i+1}$, es decir, cuando el operador Ly_i sea autoconjugado, la solución del problema (4) puede ser estimada con ayuda de las fórmulas de factorización derecha. Es más conveniente escribir la ecuación (4) en la forma

$$\Lambda y_i = (a y_{\bar{x}})_{x,i} - d_i y_i = -\varphi_i,$$
$$i = 1, \ldots, N-1,\ y_0 = 0,\ y_N = 0, \qquad (13)$$
$$a_i > 0, \quad d_i > 0.$$

Escribámosla en la forma habitual:

$$a_i y_{i-1} - c_i y_i + a_{i+1} y_{i+1} = -h^2 \varphi_i, \quad y_0 = y_N = 0,$$
$$c_i = a_i + a_{i+1} + h^2 d_i, \quad a_i > 0, \quad i = 1, 2, \ldots, N-1.$$

Veamos las fórmulas de factorización

$$y_i = \alpha_{i+1} y_{i+1} + \beta_{i+1}, \quad y_N = 0, \quad i = 1, 2, \ldots, N-1,$$
$$\alpha_{i+1} = \frac{a_{i+1}}{c_i - a_i \alpha_i}, \quad \alpha_1 = 0, \quad i = 1, 2, \ldots, N-1,$$
$$\beta_{i+1} = \frac{a_i \beta_i + \varphi_i h^2}{c_i - a_i \alpha_i}, \quad \beta_1 = 0 \quad i = 1, 2, \ldots, N-1$$

Bajo las condiciones (7) tenemos $|\alpha_{i+1}| \leqslant 1$, y ,

$$|y_i| \leqslant |y_{i+1}| + |\beta_{i+1}| \leqslant |y_N| + \sum_{s=i+1}^{N} |\beta_s| = \sum_{s=i+1}^{N} |\beta_s|.$$

Al introducir la función $a_i\beta_i = \eta_i$, obtenemos

$$\eta_{i+1} = (\eta_i + h^2\varphi_i)\,\alpha_{i+1},$$
$$|\eta_{i+1}| \leqslant |\eta_i| + h^2 |\varphi_i| \leqslant |\eta_i| + \sum_{k=1}^{i} h^2 |\varphi_k|,$$

de modo que

$$|\beta_{s+1}| \leqslant \frac{1}{a_{s+1}} h \sum_{k=1}^{s} h |\varphi_k|.$$

De resultas se obtiene para la solución del problema una estimación apriorística

$$\|y\|_c \leqslant \sum_{s=1}^{N} h \frac{1}{a_s} \sum_{k=1}^{s} h |\varphi_k| \leqslant \frac{1}{c_1} \sum_{s=1}^{N} h \sum_{k=1}^{s} h |\varphi_k|$$

para $a_s \geqslant c_1 > 0$.

Esta estimación nos será útil al estudiar la convergencia de los esquemas de diferencias.

Capítulo II

Interpolación e integración numérica

§ 1. Interpolación y aproximación de las funciones

1. Planteamiento del problema. Uno de los problemas fundamentales del análisis numérico es la interpolación de las funciones. Se requiere a menudo restablecer la función $f(x)$ para todos los valores de x en el segmento $a \leqslant x \leqslant b$, si están conocidos sus valores en cierto número finito de puntos del segmento mencionado. Dichos valores pueden ser determinados como resultado de las mediciones (observaciones) en un experimento natural, o bien como resultado de los cálculos. Además, puede ocurrir que la función $f(x)$ viene definida por cierta fórmula y el cálculo de sus valores, rigiéndose por dicha fórmula, es muy engorroso, razón por la cual resulta deseable tener para la función otra fórmula, más simple, (menos engorrosa para los cálculos) que permitiría hallar valores aproximados de la función en consideración con una exactitud necesaria en cualquier punto del segmento. De resultas, surge el siguiente problema matemático.

Supongamos que en el segmento $a \leqslant x \leqslant b$ viene prefijada una red $\overline{\omega} = \{x_0 = a < x_1 < \ldots < x_n = b\}$ y en los nodos de la red están definidos los valores de la función $y(x)$ iguales a $y(x_0) = y_0, \ldots, y(x_i) = y_i, \ldots, y(x_n) = y_n$. Se pide construir una *interpolante*, esto es, una función $f(x)$ que coincida con la función $y(x)$ en los nodos de la red:

$$f(x_i) = y_i, \qquad i = 0, 1, \ldots, n. \tag{1}$$

El objetivo principal de la interpolación es obtener un algoritmo rápido (económico) para calcular los valores de $f(x)$ en aquellos puntos x que no están contenidos en la tabla de los datos.

La cuestión principal es: cómo elegir la interpolante $f(x)$ y cómo estimar el error $y(x) - f(x)$. Las funciones interpoladoras $f(x)$ se construyen, como regla, en forma de las combinaciones lineales de ciertas funciones elementales:

$$f(x) = \sum_{k=0}^{n} c_k \Phi_k(x),$$

donde $\{\Phi_k(x)\}$ son funciones linealmente independientes fijas; $c_0, c_1, \ldots, c_n$, unos coeficientes hasta ahora desconocidos.

De las condiciones (1) obtenemos un sistema de $n+1$ ecuaciones respecto de los coeficientes $\{c_k\}$:

$$\sum_{k=0}^{n} c_k \Phi_k(x_i) = y_i, \qquad i = 0, 1, \ldots, n.$$

Supongamos que el sistema de funciones $\Phi_k(x)$ es de tal índole que, cualquiera que sea la elección de los nodos $a = x_0 < x_i < \ldots < x_n = b$, queda distinto de cero el determinante del sistema

$$\Delta(\Phi) = \begin{vmatrix} \Phi_0(x_0) & \Phi_1(x_0) & \ldots & \Phi_n(x_0) \\ \Phi_0(x_1) & \Phi_1(x_1) & \ldots & \Phi_n(x_1) \\ \cdot & \cdot & \cdot & \cdot \\ \Phi_0(x_n) & \Phi_1(x_n) & \ldots & \Phi_n(x_n) \end{vmatrix}.$$

En este caso los coeficientes c_k $(k = 0, 1, \ldots, n)$ se determinan unívocamente según las y_i $(i = 0, 1, \ldots, n)$ prefijadas.

A título de sistema de las funciones linealmente independientes $\{\Phi_k(x)\}$ se eligen más a menudo: funciones potenciales $\Phi_k(x) = x^k$ (en este caso $f = P_n(x)$ es un polinomio de grado n); funciones trigonométricas $\{\Phi_k(x) = \cos kx, \text{sen}\, kx\}$ (f es un polinomio trigonométrico). Se emplean también funciones racionales

$$\frac{\alpha_0 + \alpha_1 x + \ldots + \alpha_m x^m}{\beta_0 + \beta_1 x + \ldots + \beta_p x^p}$$

y otros tipos de funciones interpoladoras. Examinaremos aquí los polinomios de interpolación y la spline-interpolación: un caso de interpolación polinomial a trozos.

2. Interpolación polinomial. Se conoce que cualquier función $f(x)$, continua en el segmento $[a, b]$, puede ser bien aproximada mediante un polinomio $P_n(x)$ *):

TEOREMA DE WEIERSTRASS. *Para todo* $\varepsilon > 0$ *existe tal polinomio* $P_n(x)$ de grado $n = n(\varepsilon)$ que

$$\max_{x\in[a,b]} |f(x) - P_n(x)| < \varepsilon.$$

Sin embargo, este teorema no nos da la respuesta a la pregunta sobre la existencia de buen polinomio interpolador para el conjunto dado de puntos $\{(x_i, y_i)\}$.

Así pues, buscaremos el *polinomio de interpolación* en la forma

$$P_n(x) = \sum_{k=0}^{n} c_k x^k, \tag{2}$$

donde c_k son los coeficientes indeterminados. Suponiendo $f(x_i) = y_i$, obtenemos un sistema de ecuaciones lineales

$$\begin{array}{l} c_0 + c_1 x_0 + \ldots + c_n x_0^n = y_0, \\ c_0 + c_1 x_1 + \ldots + c_n x_1^n = y_1, \\ \cdots\cdots\cdots\cdots\cdots\cdots \\ c_0 + c_1 x_1 + \ldots + c_n x_n^n = y_n. \end{array}$$

El determinante de este sistema será determinante de Vandermonde distinto de cero:

$$\Delta = \begin{vmatrix} 1 & x_0 & x_0^2 & \ldots & x_0^n \\ 1 & x_1 & x_1^2 & \ldots & x_1^n \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_n & x_n^2 & \ldots & x_n^n \end{vmatrix} = \prod_{n\geqslant k>m\geqslant 0} (x_k - x_m) \neq 0.$$

De aquí proviene que el polinomio de interpolación (2) existe y es único (existe una variedad de formas para su inscripción).

A título de base de $\{\Phi_k(x)\}$ hemos elegido una base compuesta por los monomios $1, x, x^2, \ldots, x^n$. Para los cálculos resulta más cómoda la base de los polinomios de Lagrange $\{l_k(x)\}$ de grado n o bien de los *coeficientes de*

*) Véase, por ejemplo, V. A. Ilyin, E. Y. Poznyak, "Fundamentals of Mathematical Analysis" Editorial Mir, Moscú, 1982.

Lagrange:

$$l_k(x_i) = \begin{cases} 1, \text{ si } i = k, \\ 0, \text{ si } i \neq k, \; i, \; k = 0, 1, \ldots, n. \end{cases}$$

No es difícil ver que el polinomio de grado n

$$l_k(x) = l_k^{(n)}(x) =$$

$$= \frac{(x-x_0)(x-x_1) \ldots (x-x_{k-1})(x-x_{k+1}) \ldots (x-x_n)}{(x_k-x_0)(x_k-x_1) \ldots (x_k-x_{k-1})(x_k-x_{k+1}) \ldots (x_k-x_n)}$$

satisface estas condiciones. El polinomio $l_k(x)$ se define, evidentemente, del modo unívoco. Efectivamente, supongamos que existe un polinomio más $\overline{l}_k(x)$; entonces la diferencia entre ellos $\overline{l}_k(x) - l_k(x) = q_n(x)$ es un polinomio de grado n que se reduce a cero en $n + 1$ puntos x_i ($i = 0, 1, \ldots$ $\ldots, n$). Esto será posible sólo cuando $\overline{l}_k(x) - l_k(x) \equiv 0$.

El polinomio $l_k(x)\, y_k$ toma el valor y_k en el punto x_k y es nulo en todos los demás nodos x_j para $j \neq k$. De aquí se desprende que el polinomio de interpolación

$$P_n(x) = \sum_{k=0}^{n} l_k(x)\, y_k = \sum_{k=0}^{n} y_k \prod_{i \neq k} \frac{x - x_i}{x_k - x_i} \qquad (3)$$

tiene el grado no superior a n y $P_n(x_i) = y_i$. La *fórmula* (3) lleva el nombre *de Lagrange*. El número de operaciones aritméticas para el cálculo según (3) es proporcional a n^2. Para estimar la proximidad del polinomio $P_n(x)$ a la función $f(x)$ se supone que existe la $n + 1$-ésima derivada continua $f^{(n+1)}(x)$. En este caso resulta verídica la fórmula siguiente para el error:

$$f(x) - P_n(x) = \frac{f^{(n+1)}(\xi)}{(n+1)!} \prod_{j=1}^{n+1} (x - x_j), \quad \xi \in [a, b].$$

3. Fórmula de interpolación de Newton. Al emplear en los cálculos los ordenadores, es cómoda la *fórmula de interpolación* de Newton. Con el fin de escribirla se debe introducir las así llamadas diferencias divididas:

— *la diferencia dividida de primer orden*: $y(x_i, x_j) =$ $= [y(x_i) - y(x_j)]/(x_i - x_j)$;

— *la diferencia dividida de segundo orden*: $y(x_i, x_j, x_k) = [y(x_i, x_j) - y(x_j, x_k)]/(x_i - x_k)$, etc. Si $y(x) = P_n(x)$ es un polinomio de grado n, entonces para él la primera diferencia dividida $P(x, x_0) = [P(x) - P(x_0)]/(x - x_0)$ será un polinomio de grado $n - 1$; la segunda diferencia $P(x, x_0, x_1)$, polinomio de grado $n - 2$, etc., de suerte que la $(n + 1)$-ésima diferencia dividida es igual a cero.

De la definición de diferencias divididas se deduce:

$$P(x) = P(x_0) + (x - x_0) P(x, x_0),$$

$$P(x, x_0) = P(x_0, x_1) + (x - x_1) P(x, x_0, x_1),$$

$$P(x, x_0, x_1) = P(x_0, x_1, x_2) + (x - x_2) P(x, x_0, x_1, x_2),$$

etc. De aquí obtenemos para $P(x)$ la fórmula

$$P(x) = P(x_0) + (x - x_0) P(x_0, x_1) + (x - x_0)(x - x_1) \times \\ \times P(x_0, x_1, x_2) + \ldots + (x - x_0)(x - x_1) \ldots \\ \ldots (x - x_n) P(x_0, x_1, \ldots, x_n). \quad (4)$$

Si $P(x)$ es el polinomio de interpolación para la función $y(x)$, sus valores en los nodos de las redes coincidirán con los valores de la función $y(x)$, y, por consiguiente, coincidirán también las diferencias divididas. Por eso, en lugar de (4) podemos escribir

$$f(x) = y_0 + \sum_{k=1}^{n} (x - x_0)(x - x_1) \ldots \\ \ldots (x - x_{k-1}) y(x_0, x_1, \ldots, x_k)$$

(*polinomio de Newton*). Calculadas las diferencias divididas, el polinomio de Newton se calculará con toda la comodidad según el esquema de Horner

$$f(x) = y(x_0) + (x - x_0)[y(x_0, x_1) + (x - x_1) \times \\ \times [y(x_0, x_1, x_2) + \ldots]].$$

El cálculo de $f(x)$ para cada x requiere n multiplicaciones y $2n$ operaciones de sumación y sustracción.

Existen también otras fórmulas de interpolación. Entre ellas resulta más aplicable la *interpolación hermitiana*. Aquí el problema se plantea del modo siguiente. Están prefijados n nodos $\{x_i\}$, n valores de la función $\{y_i\}$ y n valores de la

derivada $\{y'_i\}$; se pide hallar tal polinomio de grado máximo $2n-1$ que se verifique

$$P(x_i) = y_i, \quad P'(x_i) = y'_i, \quad i = 1, 2, \ldots, n.$$

Si todos los x_i son distintos, existe la única solución que se halla por un método análogo al de Lagrange.

Se debe tener en cuenta que la aplicación de un polinomio de alto grado puede conducir a los problemas difíciles relacionados con los errores de redondeo.

4. Spline-interpolación. Estudiemos un caso especial de la interpolación polinomial a trozos cuando entre cualesquiera nodos vecinos de la red la función viene interpolada por un polinomio cúbico (*spline-interpolación cúbica*). Los coeficientes de dicho polinomio se determinan en cada intervalo partiendo de las condiciones de conjugación en los nodos:

$$\begin{aligned} f_i &= y_i \\ f'(x_i - 0) &= f'(x_i + 0), \\ f''(x_i - 0) &= f''(x_i + 0), \quad i = 1, 2, \ldots, n-1. \end{aligned}$$

Además, en la frontera se ponen las condiciones para $x = x_0$ y $x = x_n$

$$f''(x_0) = 0, \qquad f''(x_n) = 0. \tag{5}$$

Buscaremos el polinomio cúbico en la forma

$$f(x) = a_i + b_i(x - x_{i-1}) + c_i(x - x_{i-1})^2 + d_i(x - x_{i-1})^3, \quad x_{i-1} \leqslant x \leqslant x_i. \tag{6}$$

De la condición $f_i = y_i$ tenemos

$$\begin{aligned} f(x_{i-1}) &= a_i = y_{i-1}, \\ f(x_i) &= a_i + b_i h_i + c_i h_i^2 + d_i h_i^3 = y_i, \\ h_i &= x_i - x_{i-1}, \quad i = 1, 2, \ldots, n-1. \end{aligned} \tag{7}$$

Calculemos las derivadas:

$$\begin{aligned} f'(x) &= b_i + 2c_i(x - x_{i-1}) + 3d_i(x - x_{i-1})^2, \\ f''(x) &= 2c_i + 6d_i(x - x_{i-1}), \quad x_{i-1} \leqslant x \leqslant x_i \end{aligned}$$

y exijamos su continuidad para $x = x_i$:

$$b_{i+1} = b_i + 2c_ih_i + 3d_ih_i^2, \tag{8}$$
$$c_{i+1} = c_i + 3d_ih_i, \quad i = 1, 2, \ldots, n-1.$$

El número total de los coeficientes desconocidos es igual, evidentemente a $4n$, el número de ecuaciones (7) y (8) equivale a $4n - 2$. Dos ecuaciones que faltan las obtenemos de las condiciones (5) para $x = x_0$ y $x = x_n$:

$$c_1 = 0, \qquad c_n + 3d_nh_n = 0.$$

Expresando a base de (8) $d_i = (c_{i+1} - c_i)/3h_i$, sustituyendo esta expresión en (7) y excluyendo $a_i = y_{i-1}$, obtendremos

$$b_i = [(y_i - y_{i-1})/h_i] - \frac{1}{3}h_i(c_{i+1} + 2c_i), \quad i = 1, 2, \ldots, n-1,$$
$$b_n = [(y_n - y_{n-1})/h_n] - \frac{2}{3}h_nc_n.$$

Ahora, al sustituir las expresiones para b_i, b_{i+1} y d_i en la primera fórmula de (8), obtenemos, después de algunas transformaciones no complejas, una ecuación en diferencias de segundo orden

$$h_ic_i + 2(h_i + h_{i+1})c_{i+1} + h_{i+1}c_{i+2} = 3\left(\frac{y_{i+1} - y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right),$$
$$i = 1, 2, \ldots, n-1, \tag{9}$$

con las condiciones de contorno

$$c_1 = 0, \qquad c_{n+1} = 0, \tag{10}$$

y esta ecuación se usa para determinar c_i. La condición $c_{n+1} = 0$ es equivalente a la condición $c_n + 3d_nh_n = 0$ y a la ecuación $c_{i+1} = c_i + d_ih_i$. La ecuación en diferencias (9) con las condiciones (10) se resuelve por el método de factorización.

Se puede introducir la noción de *spline de orden m* como función que es un polinomio de grado m en cada uno de los segmentos de la red y que en todos los nodos interiores de la red satisface las condiciones de continuidad de la función y de las derivadas hasta el orden $m - 1$ inclusive. Habitualmente se usan para la interpolación los casos de $m = 3$

(spline cúbico analizado más arriba) y de $m = 1$ (*spline lineal*, correspondiente a la aproximación de la gráfica de la función $y(x)$ por una quebrada que pasa a través de los puntos (x_i, y_i)).

5. Aplicación de la interpolación. La interpolación se aplica en varios problemas relacionados con los cálculos. Indiquemos aquí algunos de estos problemas.

La elaboración de un experimento físico consistente en la construcción de las fórmulas aproximadas para las magnitudes características según datos tabulares obtenidos en los experimentos.

La construcción de las fórmulas aproximadas a base de los datos de un experimento de cálculo. En este caso surgen problemas no típicos de interpolación, ya que, corrientemente, se escriben fórmulas cuya estructura sea cuanto más simple.

La subtabulación, o sea, el espesamiento de las tablas se usa en aquellos casos cuando el cálculo inmediato de las funciones resulta difícil, o cuando se tienen pocos datos experimentales. A la máquina electrónica se introduce una tabla pequeña, mientras que los valores de la función indispensables en los cálculos se hallan, cuando sea necesario, según la fórmula de interpolación.

La interpolación se aplica también en el problema de *interpolación inversa*: está dada la tabla $y_i = y(x_i)$; se pide hallar x_i como función de y_i. A título de ejemplo de interpolación inversa puede servir el problema de búsqueda de las raíces de una ecuación.

Las fórmulas de interpolación se emplean también al calcular integrales y al escribir aproximaciones de diferencias para las ecuaciones diferenciales a base de las identidades integrales. El aparato matemático de cualquier ordenador contiene programas estándar de interpolación.

6. Aproximación media cuadrática. Hasta ahora hemos analizado la construcción de los polinomios de interpolación $y(x)$ que coinciden con los valores de la función de partida $f(x)$ en cierto conjunto de nodos sobre la red ω:

$$y(x_i) = f(x_i), \qquad x_i \in \omega.$$

La función $y(x)$ aproxima la función $f(x)$ en el intervalo de la red.

Sea $L_2[a, b]$ un espacio de funciones reales con producto escalar

$$(f, \varphi)=\int_a^b f(x)\varphi(x)\,dx$$

y norma

$$\|f\|_{L_2}=\sqrt{(f, f)}.$$

Examinemos el problema general sobre la aproximación de las funciones $f(x)$ mediante las funciones pertenecientes a L_2, sustituyendo la exigencia $y_i = f_i$ por la condición del mínimo de la norma: $\|f - y\|_{L_2}$ o de la pequeñez de la misma: $\|f - y\|_{L_2} < \varepsilon$, donde $\varepsilon > 0$ es la exactitud prefijada.

La búsqueda de $\inf \|f - y\|_{L_2}$ es el problema de encontrar la *mejor aproximación media cuadrática*. A título de $y(x)$ tomemos el *polinomio generalizado*

$$y(x)=\sum_{k=0}^{n} c_k\varphi_k(x),$$

donde $\{\varphi_k(x)\}$ es una familia de funciones ortonormalizadas en $[a, b]$

$$(\varphi_k, \varphi_m)=\delta_{km}, \qquad \delta_{km}=\begin{cases}1, & k=m,\\ 0, & k\neq m,\end{cases}$$

y c_k son unos coeficientes arbitrarios. Entonces, el problema de encontrar la mejor aproximación media cuadrática se reduce a la búsqueda del mínimo de la función de $n + 1$ variables $c_0, c_1, \ldots, c_n$:

$$\min_{(c_k)}\left\|f(x)-\sum_{k=0}^{n} c_k\varphi_k(x)\right\|=F(c_0, c_1, \ldots, c_n).$$

Calculemos la *desviación media cuadrática*

$$\|f - y\|^2 = \|f\|^2 - 2(f, y) + \|y\|^2.$$

Sustituyendo aquí las expresiones

$$(f, y)=\sum_{k=0}^{n} c_k(f, \varphi_k)=\sum_{k=0}^{n} c_k f_k, \quad f_k=(f, \varphi_k),$$

$$\|y\|^2=\sum_{k=0}^{n} c_k^2,$$

obtendremos

$$\| f-y \|^2 = \| f \|^2 + \sum_{k=0}^{n} (c_k - f_k)^2 - \sum_{k=0}^{n} f_k^2.$$

De aquí se ve que el mínimo del error se consigue para $c_k = f_k$, es decir, en la función

$$\bar{y}(x) = \bar{y}_n(x) = \sum_{k=0}^{n} f_k \varphi_k(x).$$

En este caso

$$\| f - \bar{y}_n(x) \|^2 = \| f \|^2 - \sum_{k=0}^{n} f_k^2. \tag{11}$$

De este modo, la mejor aproximación media cuadrática existe y es única. Ella lleva al problema sobre el cálculo de las integrales para determinar $c_k = f_k = (f, \varphi_k)$.

Si las funciones $\{\varphi_k\}$ forman un sistema ortonormalizado completo, se verifica la *igualdad de Parseval—Steklov*

$$\sum_{k=0}^{\infty} f_k^2 = \int_a^b f^2(x)\,dx = \| f \|^2. \tag{12}$$

Al comparar (11) con (12), encontramos

$$\| f - \bar{y}_n \|^2 = \sum_{k=n+1}^{\infty} f_k^2,$$

es decir, $\| f - \bar{y}_n \| \to 0$ cuando $n \to \infty$; la mejor aproximación media cuadrática converge hacia $f(x)$ y queda posible la aproximación con cualquier exactitud: $\| f - \bar{y}_n \| \leqslant \varepsilon$, siempre que $n \geqslant N(\varepsilon)$ (n es bastante grande).

OBSERVACIÓN. Todos los razonamientos están en vigor, si el producto escalar se toma con el peso $\rho(x) > 0$:

$$(f, \varphi) = \int_a^b f(x)\,\varphi(x)\,\rho(x)\,dx.$$

Son posibles también otros criterios de la aproximación, cuando la desviación $f - y$ se minimiza en otra norma, por ejemplo, en la norma del espacio C (*aproximación uniforme*).

Realizándose la mejor aproximación uniforme, nosotros buscamos la función $y(x)$ en la cual se consigue

$$\min_{\{y\}} \max_{a \leqslant x \leqslant b} | f(x) - y(x) |.$$

Sin embargo, hasta ahora no se ha encontrado un método que permita encontrar los coeficientes de la mejor aproximación uniforme (por el número finito de operaciones) para una función, definida en el segmento $[a, b]$. Se pueden indicar, además, otros planteamientos de los problemas de aproximación: en un conjunto discreto, en una totalidad de segmentos y otros. Se estudian también los métodos de aproximación no lineal, por ejemplo, con ayuda de las funciones racionales. Lo último resulta efectivo al elaborar los resultados de los experimentos.

§ 2. Integración numérica

1. Planteamiento del problema. Fórmulas de integración numérica (de cuadratura) más simples. El objetivo de la *integración numérica* es hallar el valor aproximado de la integral

$$J[f] = \int_a^b f(x)\,dx, \tag{1}$$

donde $f(x)$ es una función prefijada. En el segmento $[a, b]$ se introduce una red $\overline{\omega} = \{x_i: x_0 = a < x_1 < \ldots < x_i < < x_{i+1} \ldots x_N = b\}$ y como el valor aproximado de la integral se considera el número

$$J_N[f] = \sum_{i=0}^{N} c_i f(x_i), \tag{2}$$

donde $f(x_i)$ son los valores de la función $f(x)$ en los *nodos* $x = x_i$ y c_i, *factores ponderales (de peso)* que sólo dependen de los nodos, pero no son dependientes de la elección de $f(x)$. La *fórmula* (2) se denomina de *cuadratura*, o de integración numérica.

El objeto de la integración numérica con ayuda de cuadraturas consiste en búsqueda de tales nodos $\{x_i\}$ y pesos $\{c_i\}$

que el *error de la fórmula de cuadratura*

$$D[f]=\sum_{i=0}^{N} c_i f(x_i)-\int_a^b f(x)\,dx=J_N[f]-J[f]$$

sea mínimo para las funciones pertenecientes a la clase dada (la magnitud de $D[f]$ depende del grado de suavidad de $f(x)$). Al construir la fórmula de cuadratura, la integral (1) se representa, corrientemente, en forma de una suma de las integrales del tipo

$$\int_\alpha^\beta f(x)\,dx,$$

cada una de las cuales se reduce a la integral estándar por el segmento de longitud unidad:

$$L[f]=\int_0^1 f(s)\,ds \tag{3}$$

mediante una sustitución

$$x=\alpha+(\beta-\alpha)s, \tag{4}$$

$$f(x)=f(\alpha+(\beta-\alpha)s)=\bar{f}(s), \tag{5}$$

de modo que

$$\int_\alpha^\beta f(x)\,dx=\varkappa\int_0^1 \bar{f}(s)\,ds=\varkappa L[\bar{f}],\qquad \varkappa=\beta-\alpha$$

(la raya por arriba de $f(s)$ se omitirá). Convengamos en considerar que $\bar{\omega}$ es una red uniforme. En este caso podemos escribir

$$J[f]=\sum_{i=1}^{N} J_i,$$

$$J_i=\int_{x_{i-1}}^{x_i} f(x)\,dx=h\int_0^1 f(x_{i-1}+hs)\,ds.$$

Si $N = 2i_0$ es un número par, tenemos

$$J[f] = \sum_{i=1}^{i_0} J_{2i-1},$$

$$J_{2i-1} = \int_{x_{2i-2}}^{x_{2i}} f(x)\,dx = 2h\int_0^1 f(x_{2i-2} + 2hs)\,ds,$$

etc.

Así pues, el problema se reduce a la construcción de la fórmula de cuadratura para la integral (3) que se calcula por un segmento unidad. Escojamos en el segmento $0 \leqslant s \leqslant \leqslant 1$ los nodos $0 \leqslant s_0 < s_1 < \ldots < s_m \leqslant 1$ (*molde* de la fórmula de cuadratura) y a la integral (3) le pondremos en correspondencia la fórmula

$$\Lambda(f) = \sum_{k=0}^{m} p_k f(s_k). \tag{6}$$

Veamos las fórmulas de cuadratura más simples:

— *fórmula del rectángulo* (el molde contiene un nodo):

$$m = 0, \quad p_0 = 1, \quad s_0 = \frac{1}{2}, \quad \Lambda(f) = f\left(\frac{1}{2}\right);$$

— *fórmula del trapecio* (dos nodos):

$$m = 1, \quad p_0 = \frac{1}{2}, \quad p_1 = \frac{1}{2}; \quad s_0 = 0, \quad s_1 = 1$$

$$\Lambda(f) = \frac{1}{2}(f(0) + f(1));$$

— *fórmula de Simpson* (tres nodos):

$$m = 2, \quad p_0 = p_2 = \frac{1}{6}, \quad p_1 = \frac{4}{6}, \quad s_0 = 0, \quad s_1 = \frac{1}{2}, \quad s_2 = 1,$$

$$\Lambda(f) = \frac{1}{6}\left(f(0) + 4f\left(\frac{1}{2}\right) + f(1)\right)$$

y otras. En la práctica se emplean, como regla, las fórmulas con un número pequeño de nodos del molde.

Escribamos ahora las fórmulas correspondientes para la integral (1) en una red uniforme $\{x_i = ih\}$ de paso h. Teniendo presente la sustitución (4) y (5), obtendremos

— fórmula del rectángulo:

$$J_N[f] = \sum_{i=0}^{N-1} hf(x_{i+1/2}), \quad x_{i+1/2} = x_i = \frac{1}{2}h; \qquad (7)$$

— fórmula del trapecio:

$$J_N[f] = \sum_{i=0}^{N} c_i f(x_i) h, \quad c_0 = c_N = \frac{1}{2}, \quad c_i = 1,$$

$$i = 1, 2, \ldots, N-1; \qquad (8)$$

— fórmula de Simpson:

$$J_N[f] = \sum_{i=0}^{N} c_i f(x_i) \bar{h} = \frac{h}{3}(f_0 + 4f_1 + 2f_2 + 4f_3 + \ldots$$

$$\ldots + 2f_{N-2} + 4f_{N-1} + f_N) \text{ para } N = 2i_0. \qquad (9)$$

2. Construcción de las fórmulas de cuadratura. En virtud de lo dicho más arriba, exponer el problema será suficiente para la integral tipo (3), a la que se le pone en correspondencia la fórmula de cuadratura

$$\int_0^1 f(s)\,ds \approx \sum_{h=0}^{m} p_k f(s_k). \qquad (10)$$

En el caso general los nodos y los pesos son desconocidos y han de ser determinados.

Examinemos al principio un caso en que los nodos están prefijados y se requiere hallar los pesos de la fórmula de cuadratura $\{p_k\}$. Hagamos uso del requisito: la fórmula (10) debe ser exacta para cualquier polinomio $P_r(s)$ de grado $r \leqslant m$:

$$\Lambda[P_r] = L[P_r], \qquad r \leqslant m. \qquad (11)$$

Para que un polinomio de grado r satisfaga (11), basta por exigir que la fórmula de cuadratura sea exacta para cualquier monomio s^σ de grado σ ($\sigma = 0, 1, \ldots, r$). Teniendo presente que $L[s^\sigma] = 1/(\sigma + 1)$, obtenemos de (11) $m + 1$

ecuaciones

$$p_0+p_1+\ldots+p_m=1,$$
$$p_0s_0+p_1s_1+\ldots+p_ms_m=1/2,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$p_0s_0^\sigma+p_1s_1^\sigma+\ldots+p_ms_m^\sigma=1/(\sigma+1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots$$
$$p_0s_0^m+p_1s_1^m+\ldots+p_ms_m^m=1/(m+1).$$

Este sistema tiene una solución única, puesto que su determinante es el de Vandermonde, distinto de cero, si no hay nodos coincidentes, $s_0 < s_1 < \ldots s_m$.

Así que, suponiendo $m = 2$, $s_0 = 0$, $s_1 = 1/2$, $s_2 = 1$, tenemos un sistema $p_0 + p_1 + p_2 = 1$, $p_1/2 + p_2 = 1/2$, $p_1/4 + p_2 = 1/3$, cuya solución está representada por los pesos de la fórmula de Simpson: $p_0 = p_2 = 1/6$, $p_1 = 4/6$. De este modo, la fórmula de Simpson es exacta para un polinomio de segundo grado. No obstante, por ser simétrica, será también exacta para todos los polinomios de tercer grado:

$P_3(s) = 1 + \alpha_1(s - 1/2) + \alpha_2(s - 1/2)^2 + \alpha_3(s - 1/2)^3$, puesto que es exacta para $f(s) = (s - 1/2)^3$. En efecto,

$$\Lambda\left[\left(s-\frac{1}{2}\right)^3\right]=\frac{1}{6}\left(\left(-\frac{1}{2}\right)^3+4\cdot 0+\left(\frac{1}{2}\right)^3\right)=0,$$

$$L\left[\left(s-\frac{1}{2}\right)^3\right]=\int_0^1\left(s-\frac{1}{2}\right)^3 ds=0.$$

Las fórmulas del rectángulo y de trapecio son exactas para una función lineal, es decir, para un polinomio de primer grado, de lo que es fácil convencerse inmediatamente.

En el caso general, a título de $P_m(s)$ puede elegirse el polinomio de interpolación de Lagrange

$$P_m(s)=\sum_{k=0}^{m} l_k^{(m)}(s)\, f(s_k),$$

donde $l_k^{(m)}(s)$ es el coeficiente interpolador de Lagrange. De la igualdad

$$L[P_m]=\int_0^1 P_m(s)\,ds=\sum_{k=0}^{m} f(s_k)=\int_0^1 l_k^{(m)}(s)\,ds=\sum_{k=0}^{m} p_k f(s_k)$$

se ve que la fórmula (10) es exacta para un polinomio de grado m, si los factores ponderales p_k se determinan según la fórmula

$$p_k = \int_0^1 l_k^{(m)}(s)\, ds. \qquad (12)$$

Las fórmulas de este tipo se llaman *fórmulas* de cuadratura *de Cotes*.

Aduzcamos como ejemplos de las fórmulas de cuadratura dos fórmulas más:

en el molde tetrapuntual, $s_k = k/3$ $(k = 0, 1, 2, 3)$, $m = 3$:

$$\Lambda(f) = \frac{1}{8}\left(f(0) + 3f\left(\frac{1}{3}\right) + 3f\left(\frac{2}{3}\right) + f(1)\right),$$

$$p_0 = p_3 = \frac{1}{8}, \quad p_1 = p_2 = \frac{3}{8},$$

en el molde pentapuntual, $s_k = k/4$ $(k = 0, 1, 2, 3, 4)$, $m = 4$:

$$\Lambda(f) = \frac{1}{90}\left(7f(0) + 32f\left(\frac{1}{4}\right) + 12f\left(\frac{1}{2}\right) + \right.$$

$$\left. + 32f\left(\frac{3}{4}\right) + 7f(1)\right),$$

$$p_0 = p_4 = \frac{7}{90}, \quad p_1 = p_3 = \frac{32}{90}, \quad p_2 = \frac{12}{90}.$$

Los moldes de las cinco fórmulas de cuadratura aducidas más arriba constan de los nodos simétricos con relación al centro $s = 1/2$ del segmento $0 \leqslant s \leqslant 1$.

3. Fórmula de Taylor con término residual en la forma integral. Al investigar el error de la fórmula de cuadratura nos hará falta la fórmula de Taylor con término residual en la forma integral:

$$f(s) = f(0) + sf'(0) + \frac{s^2}{2} f''(0) + \ldots + \frac{s^n}{n!} f^{(n)}(0) + R_{n+1}(s), \qquad (13)$$

$$R_{n+1}(s) = \int_0^s \frac{(s-t)^n}{n!} f^{(n+1)}(t)\, dt.$$

Dicha fórmula puede ser demostrada por inducción respecto de n. Para $n = 0$ es justa:

$$f(s) = f(0) + R_1(s), \quad R_1(s) = \int_0^s f'(t)\, dt.$$

Admitamos que es justa para n. Integrando por partes, obtenemos una correlación

$$\int_0^s \frac{(s-t)^n}{n!} f^{(n+1)}(t)\, dt =$$

$$= -\frac{(s-t)^{n+1}}{(n+1)!} f^{(n+1)}(t)\Big|_0^s + \int_0^s \frac{(s-t)^{n+1}}{(n+1)!} f^{(n+2)}(t)\, dt =$$

$$= \frac{s^{n+1}}{(n+1)!} f^{(n+1)}(0) + \int_0^s \frac{(s-t)^{n+1}}{(n+1)!} f^{(n+2)}(t)\, dt, \quad (14)$$

la cual demuestra la fórmula (13) precisamente para $n + 1$. Introduciendo la función

$$K_n(\xi) = \begin{cases} \xi^n/n! & \text{para } \xi \geqslant 0, \\ 0 & \text{para } \xi < 0, \end{cases} \quad (15)$$

escribamos la fórmula para el término residual R_{n+1} en la forma

$$R_{n+1}(s) = \int_0^1 K_n(s-t) f^{(n+1)}(t)\, dt. \quad (16)$$

4. Fórmula para el error de la fórmula de cuadratura. Pasemos a la deducción de una fórmula para el error de la fórmula de cuadratura

$$\Delta(f) = \Lambda[f] - L[f] \quad (17)$$

en la clase $C^{(n+1)}$ de funciones que tienen la $(n+1)$-ésima derivada continua en el segmento $0 \leqslant s \leqslant 1$: $f(s) \in C^{(n+1)}[0, 1]$. En este caso sirve la fórmula (13) o bien

$$f(s) = P_n(s) + R_{n+1}(s), \quad P_n(s) = \sum_{\sigma=0}^{n} \frac{s^\sigma}{\sigma!} f^{(\sigma)}(0). \quad (18)$$

De lo expuesto anteriormente (véase el p. 2) está claro que para un polinomio $P_n(s)$ de grado n la fórmula (10) es exacta en dos casos: para $n \leqslant m + 1 = n_0$, si m es par y la fórmula es simétrica; para $n \leqslant m = n_0$ en todos los demás casos. Supondremos por ahora que

$$\Lambda[P_n] = L[P_n], \quad \text{es decir,} \quad n \leqslant n_0. \tag{19}$$

Volvamos ahora a la diferencia $\Delta(f)$ y sustituyamos $f = P_n + R_{n+1}$ en (17). Tomando en consideración (16) y (19), obtendremos

$$\Delta(f) = \Lambda[f] - L[f] =$$
$$= (\Lambda[P_n] - L[P_n]) + (\Lambda[R_{n+1}] - L[R_{n+1}]) =$$
$$= \Lambda[R_{n+1}] - L[R_{n+1}] = \sum_{k=0}^{m} p_k \int_0^1 K_n(s_k - t) f^{(n+1)}(t)\,dt -$$
$$- \int_0^1 \int_0^1 K_n(s - t) f^{(n+1)}(t)\,dt\,ds =$$
$$= \int_0^1 \Big[\sum_{k=0}^{m} p_k K_n(s_k - t) - \int_0^1 K_n(s - t)\,ds\Big] f^{(n+1)}(t)\,dt.$$

Haciendo uso de la expresión (15) para $K_n(s - t)$, hallamos

$$\int_0^1 K_n(s - t)\,ds = \int_t^1 \frac{(s-t)^n}{n!}\,ds = \frac{(1-t)^{n+1}}{(n+1)!}.$$

De resultas, la fórmula para el error toma la forma

$$\Delta(f) = \int_0^1 F_{n+1}(t) f^{(n+1)}(t)\,dt, \tag{20}$$

donde

$$F_{n+1}(t) = \sum_{k=0}^{m} p_k K_n(s_k - t) - \frac{(1-t)^{n+1}}{(n+1)!}. \tag{21}$$

De aquí se desprende la estimación para el error

$$|\Delta(f)| \leqslant M_{n+1} c_{n+1} \tag{22}$$

para $|f^{(n+1)}(t)| \leqslant M_{n+1}$, donde $M_{n+1} > 0$ es una constante, y para

$$c_{n+1} = \int_0^1 |F_{n+1}(t)|\, dt.$$

Si $F_{n+1}(t)$ no cambia de signo en el segmento $0 \leqslant s \leqslant 1$, *entonces, en virtud del teorema del valor medio, tenemos*

$$\Delta(f) = f^{n+1}(\xi) \int_0^1 F_{n+1}(t)\, dt, \quad \xi \in [0, 1].$$

5. **Estimación del error de las fórmulas concretas.** Nues*tro objetivo es obtener la estimación del error* $\Delta(\bar{f}) = \Lambda[\bar{f}] - L[\bar{f}]$ de la fórmula de cuadratura para la integral estándar (3). Al pasar a las fórmulas para las integrales (1) y (3), se debe tener en cuenta que

$$\frac{d^{\sigma}\bar{f}(s)}{ds^{\sigma}} = \varkappa^{\sigma} \frac{d^{\sigma}f(x)}{dx^{\sigma}},$$

$\bar{f}(s) = f(x), \quad x = \alpha + (\beta - \alpha) s, \quad dx = \varkappa\, ds, \ \varkappa = \beta - \alpha.$
Por eso, para el error

$$d[f] = \sum_{k=0}^{m} \varkappa p_k f(x_k) - \int_\alpha^\beta f(x)\, dx = \varkappa \Delta(\bar{f})$$

es justa, en virtud de (22), la fórmula

$$|d[f]| \leqslant c_{n+1} \varkappa^{n+2} \max_{x \in [\alpha, \beta]} |f^{(n+1)}(x)|,$$

$$c_{n+1} = \int_0^1 |F_{n+1}(t)|\, dt.$$

Para el cálculo del error $J_N[f] - J[f]$ es necesario, evidentemente, sumar sobre la red los errores $|D[f]|$.

Veamos las fórmulas de cuadratura más simples.

1) FORMULA DEL RECTANGULO: $m = 0$, $p_0 = 1$, $s_0 = 1/2$, $\Lambda(\bar{f}) = \bar{f}(1/2)$. Debido a la fórmula (20) tenemos

$$\Delta_1[\bar{f}] = \int_0^1 F_2(t) \bar{f}''(t)\, dt, \quad F_2(t) = K_1\left(\frac{1}{2} - t\right) - \frac{(1-t)^2}{2},$$

es decir, $F_2(t) = -(1-t)^2/2 < 0$ para $t > 1/2$, $F_2(t) = (1/2 - t) - (1-t)^2/2 = -t^2/2 < 0$ para $t < 0$, es decir, $F(t) < 0$ es una función de signo constante y

$$\Delta_1(\bar{f}) = \bar{f}''(\eta) \int_0^1 F_2(t)\,dt = -\frac{\bar{f}''(\eta)}{24}, \quad \eta \in (0, 1).$$

De aquí se infiere que

$$d_i[f] = hf(x_{i-1/2}) - \int_{x_{i-1}}^{x_i} f(x)\,dx = -\frac{h^3}{24} f''(\xi_i),$$
$$\xi_i \in [x_{i-1}, x_i]. \qquad (23)$$

Sumando según $i = 1, 2, \ldots, N$, y teniendo presente que la media aritmética es igual a

$$\sum_{i=1}^{N} hf''(\xi_i) = \frac{b-a}{N} \sum_{i=1}^{N} f''(\xi_i) = f''(\xi^*)(b-a), \ \xi^* \in [a, b],$$

obtenemos para el error la fórmula del rectángulo:

$$D_N(f) = -\frac{h^2}{24} f''(\xi^*)(b-a).$$

Si $f(x)$ tiene derivadas continuas por lo menos de cuarto orden, $f(x) \in C^{(n)}$ $(n \geqslant 4)$, podemos anotar el desarrollo asintótico para el error:

$$D_N(f) = \alpha_2 h^2 + \alpha_4 h^4, \qquad (24)$$

donde

$$\alpha_2 = -\frac{1}{24} \int_a^b f''(x)\,dx = -\frac{1}{24}[f'(b) - f'(a)].$$

En efecto, al sustituir en (20) la expresión

$$f''(t) = \bar{f}''\left(\frac{1}{2}\right) + \left(t - \frac{1}{2}\right)\bar{f}'''\left(\frac{1}{2}\right) + \frac{1}{2}\left(t - \frac{1}{2}\right)^2 \bar{f}^{\mathrm{IV}}(\eta), \quad \eta \in (0, 1)$$

hallemos, después de ciertos cálculos no complejos,

$$\Delta_1(\bar{f}) = -\frac{1}{24}\bar{f}''\left(\frac{1}{2}\right) + \frac{1}{960}\bar{f}^{\text{IV}}(\eta), \quad \eta \in (0, 1).$$

De aquí se deduce que

$$D_N(f) = -\frac{h^2}{24}\sum_{i=1}^{N} hf''_{i-1/2} + \frac{h^4}{960}\sum_{i=1}^{N} hf^{\text{IV}}(\xi_i).$$

Al tomar en consideración que, en virtud de (23),

$$\sum_{i=1}^{N} hf''_{i-1/2} = \int_a^b f''(x)\,dx - \frac{h^2}{24}f^{\text{IV}}(\xi^*)\cdot(b-a), \quad \xi^* \in [a, b],$$

obtenemos el desarrollo (24).

De (24) se ve que la fórmula del rectángulo tiene el cuarto grado de precisión: $D_N(f) = O(h^4)$, si la función $f(x)$ satisface la condición $f'(a) = f'(b)$. Si se conocen $f'(a)$ y $f'(b)$, podemos poner $f(x) = \varphi(x) + \alpha x + \beta x^2$, donde $\varphi(x)$ satisface la condición $\varphi'(a) = \varphi'(b)$, siempre que α y β se elijan del modo siguiente

$$\alpha = \frac{bf'(a) - af'(b)}{b-a}, \quad \beta = \frac{f'(b) - f'(a)}{2(b-a)}.$$

Entonces

$$\int_a^b f(x)\,dx = \int_a^b \varphi(x)\,dx + c,$$

$$c = \frac{1}{2}\alpha(b^2 - a^2) + \frac{1}{6}\beta(b^3 - a^3).$$

La integral de $\varphi(x)$ se calcula según la fórmula del rectángulo con la exactitud de $O(h^4)$.

2) FÓRMULA DEL TRAPECIO: $m = 1$, $p_0 = p_1 = 1/2$, $s_0 = 0$, $s_1 = 1$,

$$\Lambda(\bar{f}) = \frac{1}{2}(\bar{f}(0) + \bar{f}(1)).$$

La función $F_2(t) = \frac{1}{2}t(1-t) > 0$ es de signo constante, por lo cual queda válida la estimación

$$D_N(f) = \frac{h^2}{12}f''(\xi^*)\cdot(b-a), \quad \xi^* \in [a, b],$$

es decir, el coeficiente de h^2 en la expresión para el error de la fórmula del trapecio es dos veces mayor que para la fórmula del rectángulo. Reiterando los razonamientos, análogos a los citados más arriba, nos convencemos de que es justa la fórmula

$$D_N(f) = -2\alpha_2 h^2 + \alpha_4 h^4 \quad \text{para} \quad f \in C^{(n)}, \quad n \geqslant 4,$$

donde α_2 se determina de acuerdo con (24), $\alpha_4 = O(1)$.

3. FÓRMULA DE SIMPSON: $m = 2$, $s_0 = 0$, $s_1 = 1/2$, $s_2 = 1$, $p_0 = p_2 = 1/2$, $p_1 = 4/6$.

$$\Lambda(\bar{f}) = \frac{1}{6}\left(\bar{f}(0) + 4\bar{f}\left(\frac{1}{2}\right) + f(1)\right).$$

Por cuanto la fórmula de Simpson es exacta para un polinomio de tercer grado, entonces $n = 3$ y calculamos:

$$\Delta_3(\bar{f}) = \int_0^1 F_4(t)\bar{f}^{IV}(t)\,dt,$$

$$F_4(t) = \frac{1}{6}(K_3(0-t) + K_3(1-t)) + \\ + \frac{4}{6}K_3\left(\frac{1}{2}-t\right) - \frac{(1-t)^4}{24}.$$

De aquí encontramos

$$F_4(t) = \frac{1}{72}(2t^3 - 3t^4), \quad t < \frac{1}{2};$$

$$F_4(t) = \frac{1}{72}(2(1-t)^3 - 3(1-t)^4), \quad t > \frac{1}{2},$$

$$F_4(t) > 0 \text{ para todos los } t \in (0, 1),$$

y, por consiguiente,

$$\int_0^1 F_4(t)\,dt = \frac{1}{2880}$$

de suerte que es exacta la fórmula

$$\Delta_3(\bar{f}) = \frac{1}{2880}\bar{f}^{IV}(\eta), \quad \eta \in (0, 1).$$

Pasando a las integrales respecto de x y teniendo presente que $\varkappa = 2h$, $\bar{f}^{IV}(\eta) = (2h)^4 f^{IV}(\xi_1)$, obtendremos

$$D_N(f) = \sum_{i=0}^{i_0-1} 2h \left\{ \frac{f_{i-1}+4f_i+f_{i+1}}{6} - \frac{1}{2h} \int_{x_{i-1}}^{x_{i+1}} f(x)\,dx \right\} =$$

$$= \frac{b-a}{180} h^4 f^{IV}(\xi^*), \quad \xi^* \in [a, b],$$

donde $N = 2i_0$, $h = 1/N$.

Si es que $f(x) \in C^{(n)}$ $(n \geqslant 6)$, entonces podemos obtener un desarrollo de la forma

$$D_N(f) = \alpha_4 h^4 + \alpha_6 h^6, \quad \alpha_6 = O(1),$$

$$\alpha_4 = \frac{1}{180} \int_0^1 f^{IV}(x)\,dx = \frac{1}{180}(f'''(1) - f'''(0)).$$

6. Aumento del orden de exactitud. Método de Runge. Para las fórmulas de cuadratura (por la analogía con lo anterior) se puede obtener un desarrollo asintótico de la forma

$$D_N(f) = J_N(f) - J(f) = \alpha_2 h^2 + \alpha_4 h^4 + \alpha_6 h^6 + \ldots,$$

si $f(x)$ es una función suficientemente suave. En este caso $|\alpha_{k+2}|$ es considerablemente inferior a $|\alpha_k|$ $(k = 2, 4)$, razon por la cual el aumento del orden de exactitud de la fórmula de cuadratura resulta muy importante.

Realicemos los cálculos sobre dos redes uniformes con los pasos h_1 y h_2, respectivamente, y hallemos las expresiones

$$J^{h_1}[f] = J_{N_1}[f] \quad \text{y} \quad J^{h_2}[f] = J_{N_2}[f], \quad h_1 N_1 = h_2 N_2 = b - a.$$

Exijamos que el error para su combinación lineal

$$\tilde{D}^h(f) = D^{h_1}(f) + (1-\sigma) D^{h_2}(f)$$

sea una magnitud de orden superior en comparación con D^{h_1} y D^{h_2}. Si para $D^h = D_N$ tiene lugar la fórmula del tipo

$$D^h = J^h[f] - J[f] = \alpha_p h^p + \alpha_q h^q + \ldots, \quad q > p,$$

entonces para $\widetilde{D}^h = (\sigma J^{h_1}[f] + (1-\sigma) J^{h_2}[f]) - J[f]$ obtendremos

$$\widetilde{D}^h(f) = \alpha_p(\sigma h_1^p + (1-\sigma) h_2^p) + \alpha_q(\sigma h_1^q + (1-\sigma) h_2^q) + \ldots$$

Elijamos el parámetro σ, partiendo de la condición $\sigma h_1^p + (1-\sigma) h_2^p = 0$:

$$\sigma = h_2^p/(h_2^p - h_1^p).$$

En este caso tendremos

$$\widetilde{D}^h(f) = \alpha_q(\sigma h_1^q + (1-\sigma) h_2^q) + \ldots = O(h^q), \quad h = \text{máx}(h_1, h_2),$$

con la particularidad de que $\sigma h_1^q + (1-\sigma) h_2^q < 0$. Así, por ejemplo, si $p = 2$, $q = 4$, entonces $\widetilde{D}^h(f) = -\alpha_4 h_1^2 h_2^2 + \ldots = O(h^4)$. De este modo, al realizar los cálculos sobre dos redes con los pasos h_1 y $h_2 \neq h_1$, hemos aumentado el orden de exactitud en 2 (en $q - p$) para $\widetilde{J} = \sigma J^{h_1} + (1-\sigma) J^{h_2}$.

Observemos que combinando la fórmula del trapecio $J^{2h}_{\text{trap}}[f]$ y la del rectángulo $J^{2h}_{\text{rect}}[f]$, ambas con paso $2h$, obtendremos la fórmula de Simpson J^h_{Simp} con paso h:

$$J^h_{\text{Simp}}[f] = \frac{1}{3} J^{2h}_{\text{trap}}[f] + J^{2h}_{\text{rect}}[f] =$$

$$= \frac{h}{6}(f_0 + 4f_1 + 2f_2 + \ldots + 2f_{2N-2} + 4f_{2N-1} + f_{2N}),$$

donde $h = (b-a)/(2N)$.

El *método de cálculo sobre varias redes* se aplica para el aumento del orden de exactitud incluso en aquel caso cuando se desconoce el orden del término principal del error (*proceso de Aitken*). Supongamos que para el error tiene lugar la representación

$$D^h(f) = \alpha_p h^p + \alpha_q h^q + \ldots, \qquad q > p,$$

de suerte que

$$J^h[f] = J[f] + \alpha_p h^p + \alpha_q h^q + \ldots$$

Realicemos los cálculos sobre tres redes: $h_1 = h$, $h_2 = \rho h$, $h_3 = \rho^2 h$ $(0 < \rho < 1)$. Determinemos, al principio, p, me-

nospreciando el término $O(h^q)$. Formemos una razón

$$A=\frac{J^{h_1}[f]-J^{h_2}[f]}{J^{h_2}[f]-J^{h_3}[f]}\approx\frac{h_1^p-h_2^p}{h_2^p-h_3^p}=\frac{1-\rho^p}{\rho^p(1-\rho^p)}=\left(\frac{1}{\rho}\right)^p$$

y hallemos

$$p\approx\ln A/\ln\frac{1}{\rho}.$$

Sabiendo el valor aproximado de p, se puede, empleando el método de Runge expuesto más arriba, aumentar el orden de exactitud. Con este fin formemos una combinación $\widetilde{J}^h = \sigma J^{h_1} + (1-\sigma)J^{h_2}$ y elijamos σ de una manera tal que se verifique $\sigma h_1^p + (1-\sigma)h_2^p = (\sigma + (1-\sigma)\rho^p)h^p = 0$, es decir, $\sigma = \rho^p/(\rho^p - 1) = 1/(1-A)$. Entonces, para el error $\widetilde{D}^h = \widetilde{J}^h - J$ obtenemos

$$\widetilde{D}^h(f) = O(h^q).$$

Todos estos razonamientos tienen sentido, por supuesto, si la función $f(x)$ tiene una suavidad correspondiente.

7. Fórmulas de cuadratura de otro tipo. Sin perturbar la generalidad de razonamientos podemos considerar

$$J[f]=\int_0^1 f(x)\,dx. \tag{25}$$

Hasta ahora se han analizado las fórmulas de cuadratura con los nodos prefijados $\{x_k\}$:

$$J_N[f]=\sum_{k=0}^{N} c_k f(x_k). \tag{26}$$

Las fórmulas citadas son exactas para todos los polinomios de grado N. Si se consideran desconocidos no sólo c_k, sino también los nodos x_k, podemos exigir que la fórmula de cuadratura (26) sea exacta para todos los polinomios de grado $2N-1$. La fórmula de esta índole lleva el nombre de Gauss. Exigiendo que para los monomios $1, x, x^2, \ldots,$

$\ldots, x^m, \ldots, x^N$ la fórmula sea exacta, es decir, que

$$J_N[x^m] = \sum_{k=0}^{N} c_k x_k^m = \int_0^1 x^m\,dx = \frac{x^{m+1}}{m+1}\Big|_0^1 =$$
$$= \frac{1}{m+1}, \quad m = 0, 1, \ldots, 2N-1,$$

obtendremos $2N + 2$ ecuaciones para los nodos y pesos

$$c_0 + c_1 + \ldots + c_N = 1$$
$$c_0 x_0 + c_1 x_1 + \ldots + c_N x_N = 1/2,$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$c_0 x_0^m + c_1 x_1^m + \ldots + c_N x_N^m = 1/(m+1),$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$c_0 x_0^{2N+1} + c_1 x_1^{2N+1} + \ldots + c_N x_N^{2N+1} = 1/(N+1).$$

El número total de las incógnitas es igual a $2N + 2$, es decir, $N + 1$ nodos y $N + 1$ factores ponderales desconocidos. El número de ecuaciones es también igual a $2N + 2$. Se puede demostrar que el sistema escrito de ecuaciones tiene la solución.

Aduzcamos la fórmula de Gauss más sencilla para $N = 2$:

$$J_N[f] = \frac{5}{18} f(x_0) + \frac{8}{18} f(x_1) + \frac{5}{18} f(x_2),$$

donde

$$x_0 = \frac{1-\sqrt{0,6}}{2}, \quad x_2 = \frac{1+\sqrt{0,6}}{2}, \quad x_1 = \frac{1}{2}.$$

Las fórmulas de Gauss proporcionan buena precisión con el número reducido de nodos.

De un ejemplo más sirve la *fórmula* de cuadratura *de Chébishev* en la que se eligen los mejores nodos bajo la suposición de que todos los pesos son iguales. En este caso

$$J_N[f] = \frac{1}{N} \sum_{k=1}^{N} f(x_k).$$

Exigiendo que la fórmula sea exacta para $f(x) = x, x^2, \ldots$ $\ldots, x^N$, obtendremos N ecuaciones para determinar $x_1, x_2, \ldots$

$\ldots, x_N$:

$$x_1^m + x_2^m + \ldots + x_N^m = \frac{1}{m+1}, \quad m = 1, 2, \ldots, N.$$

Estas ecuaciones tienen soluciones para $m = 1, 2, \ldots, 7, 9$, y para $m = 8$ y $m \geqslant 10$ no tienen raíces reales. Cuando $m = 3$, la fórmula de Chébishev tiene por expresión

$$\int_0^1 f\,dx \approx J_3[f] = \frac{1}{3}\left[f\left(\frac{1}{2} - \frac{1}{4}\sqrt{2}\right) + f\left(\frac{1}{2}\right) + f\left(\frac{1}{2} + \frac{1}{4}\sqrt{2}\right)\right].$$

Para ella el coeficiente de $\| f^{\text{IV}} \|_C$ en la estimación del error es dos veces menor que para la fórmula de Simpson.

OBSERVACIONES. En ciertos casos al cálculo de las integrales le debe preceder su transformación, teniendo en cuenta los rasgos específicos de la función subintegral. Ejemplos:

1) $f(x) = \frac{1}{2\sqrt{x}} f_0(x)$, $f_0(0) \neq 0$, es decir, $f(x)$ tiene una singularidad cuando $x = 0$. Esta singularidad se elimina por el cambio de variable:

$$\int_0^1 f(x)\,dx = \int_0^1 \frac{f_0(x)}{2\sqrt{x}}\,dx = \int_0^1 f_0(x)\,d\sqrt{x} = \int_0^1 f(t^2)\,dt, \quad t = \sqrt{x}.$$

2) *La función subintegral tiene el carácter exponencial*: $f(x) \approx ce^{ax}$, es decir, la función $\ln f(x)$ es lineal. Representemos $f(x)$ como $f(x) = \exp\{\ln f(x)\}$, interpolemos $\ln f(x)$ linealmente en el segmento $[x_{i-1}, x_i]$:

$$\ln f(x) = \frac{x_i - x}{x_i - x_{i-1}} \ln f_{i-1} + \frac{x - x_{i-1}}{x_i - x_{i-1}} \ln f_i,$$

e integremos después respecto de x entre x_{i-1} y x_i. Esta fórmula resulta útil en los cálculos prácticos.

3) Si $f(x)$ es una *función rápidamente oscilante*, de modo que puede ser escrita en la forma $f(x) = y(x) \cos \omega x$, donde

la frecuencia $\omega \gg 1$ es grande, entonces calculando una integral se puede recurrir al procedimiento siguiente. Primero integramos por partes:

$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \int_{x_{i-1}}^{x_i} y(x)\cos\omega x\,dx =$$

$$= \frac{1}{\omega}\, y \operatorname{sen}\omega x \Big|_{x_{i-1}}^{x_i} - \frac{1}{\omega}\int_{x_{i-1}}^{x_i} y'(x)\operatorname{sen}\omega x\,dx.$$

Si $y(x)$ es lineal en $[x_{i-1}, x_i]$, entonces la integral del segundo miembro se calcula en la forma explícita. Si $y(x)$ es un polinomio de grado n, la integración por partes se realiza n veces.

Capítulo III

Resolución numérica de los sistemas de ecuaciones algebraicas lineales

En este capítulo se estudian los métodos de resolución numérica de los sistemas de ecuaciones algebraicas lineales, es decir, los métodos numéricos del álgebra lineal. Existen dos tipos de métodos: directos e iterativos. Analizamos, ante todo, el método de eliminación de Gauss para los sistemas del tipo general y las variantes: método de factorización y métodos de factorización matricial para los sistemas del tipo especial (con matrices tridiagonal o tridiagonal por bloques). Estos son los *métodos directos*. Su eficiencia depende del orden del sistema y de la estructura de la matriz.

Al estudiar los métodos *iterativos*, consideramos todo sistema de ecuaciones como una ecuación operacional de la primera especie $Au = f$, y exponemos la teoría general de los métodos iterativos para ecuaciones operacionales con suposiciones mínimas respecto del operador A. La teoría general permite demostrar la convergencia de las iteraciones para el método de Seidel y para el método de relajación superior con restricciones mínimas para el operador A. Se han analizado dos clases de los métodos: 1) para el caso en que se conocen las fronteras $\gamma_1 > 0$ y $\gamma_2 \geqslant \gamma_1$ del espectro del operador A en cierto espacio energético H_D; 2) para el caso en que las fronteras γ_1 y γ_2 no se conocen. Es de gran eficacia el método triangular alternado que se estudia en el § 5.

§ 1. *Sistemas de ecuaciones algebraicas lineales*

1. Sistemas de ecuaciones. El problema fundamental del álgebra lineal consiste en la resolución del sistema de ecuaciones

$$Au = f, \qquad (1)$$

donde $u = (u^{(1)}, \ldots, u^{(N)})$ es el vector buscado, $f = (f^{(1)}, f^{(2)}, \ldots, f^{(N)})$ es un vector conocido de dimensión N, $A = (a_{ij})$ $(i, j = 1, 2, \ldots, N)$ es una matriz cuadrada de dimensión $N \times N$ con elementos a_{ij}.

Se supondrá que la matriz A es regular, $\det A \neq 0$, de modo que la ecuación $Au = 0$ tiene sólo una solución trivial, y el sistema (1) tiene la única solución

$$u = A^{-1}f.$$

En el curso de álgebra lineal la solución del sistema (1) se expresa, corrientemente, según las fórmulas de Cramer como una razón de los determinantes. Dichas fórmulas no sirven para la resolución numérica del sistema (1), puesto que requieren el cálculo de $N + 1$ determinantes, lo que, a su vez, *exige un gran número de operaciones* aritméticas (hasta $N!$). Si incluso escogemos el mejor método, para el cálculo de un solo determinante se necesitará aproximadamente tanto tiempo que se requiere para la resolución de un sistema de ecuaciones lineales por los métodos numéricos modernos. Además, hemos de tener en cuenta, que los cálculos según las fórmulas de Cramer conducen con frecuencia a los grandes errores de redondeo.

La peculiaridad de la mayoría de los métodos numéricos para (1) consiste en que se abandona la idea de buscar la matriz inversa. El requisito principal que se levanta ante el método de resolución es el mínimo de operaciones aritméticas suficientes para la búsqueda de una solución aproximada con la precisión prefijada $\varepsilon > 0$ (eficiencia del método numérico).

2. Casos particulares de los sistemas. No es difícil resolver el sistema (1) *en los casos particulares* que van abajo. Sea A una matriz *diagonal*, es decir, $a_{ij} = 0$, $j \neq i$, $a_{ii} \neq 0$ $(i, j = 1, 2, \ldots, N)$. Entonces, el sistema tiene por expresión

$$a_{ii}u^{(i)} = f^{(i)},$$

de donde encontramos

$$u^{(i)} = f^{(i)}/a_{ii}, \quad i = 1, 2, \ldots, N.$$

Si A es la *matriz triangular inferior*, es decir, si $a_{ij} = 0$ para $j > i$ $(i, j = 1, 2, \ldots, N)$, $a_{ii} \neq 0$,

$$A = \begin{bmatrix} a_{11} & 0 & \ldots & 0 \\ a_{21} & a_{22} & \ldots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ a_{N1} & a_{N2} & \ldots & a_{NN} \end{bmatrix},$$

entonces el sistema de ecuaciones tiene por expresión

$$\begin{aligned} a_{11}u^{(1)} &= f^{(1)}, \\ a_{21}u^{(1)} + a_{22}u^{(2)} &= f^{(2)}, \\ \cdots\cdots\cdots\cdots & \cdots \\ a_{N1}u^{(1)} + a_{N2}u^{(2)} + \ldots + a_{NN}u^{(N)} &= f^{(N)}. \end{aligned}$$

Los componentes del vector u se hallan sucesivamente al pasar de n a $n+1$, partiendo de $n = 1$:

$$u^{(1)} = \frac{f^{(1)}}{a_{11}}; \qquad u^{(2)} = \frac{1}{a_{22}}\left(f^{(2)} - a_{21}u^{(1)}\right), \ \ldots,$$

$$u^{(n+1)} = \frac{1}{a_{n+1,\,n+1}}\left(f^{(n+1)} - \sum_{k=1}^{n} a_{n+1,\,k}u^{(k)}\right),$$

$$n = 2, 3, \ldots, N-1.$$

Para hallar el vector $u = (u^{(1)}, \ldots, u^{(N)})$ se necesitan $1 + 3 + 5 + \ldots + 2N - 1 = N^2$ operaciones aritméticas.

Si A es la *matriz triangular superior*, es decir, si $a_{ij} = 0$ para $j < i$, $a_{ii} \neq 0$ $(i, j = 1, 2, \ldots, N)$,

$$A = \begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1N} \\ & a_{22} & \ldots & a_{2N} \\ \cdots & \cdots & \cdots & \cdots \\ & 0 & & a_{N-1,\,N-1} + a_{N-1,N} \; a_{NN} \end{bmatrix},$$

entonces el sistema (1) tiene por expresión

$$\begin{aligned} a_{11}u^{(1)} + a_{12}u^{(2)} + \ldots + a_{1N}u^{(N)} &= f^{(1)}, \\ a_{22}u^{(1)} + \ldots + a_{2N}u^{(N)} &= f^{(2)}, \\ \cdots\cdots\cdots\cdots & \cdots \\ a_{N-1,\,N-1}u^{(N-1)} + a_{N-1,\,N}u^{(N)} &= f^{(N-1)}, \\ a_{NN}u^{(N)} &= f^{(N)}. \end{aligned}$$

Los componentes del vector $u = A^{-1}f$ se determinan sucesivamente, al pasar de $n+1$ a n, según las fórmulas

$$u^{(N)} = \frac{f^{(N)}}{a_{NN}}, \ \ldots, \ u^{(n)} = \frac{1}{a_{nn}}\left(f^{(n)} - \sum_{k=n+1}^{N} a_{nk}u^{(k)}\right), \ \ldots$$
$$n = N-1, \ N-2, \ \ldots, 2, 1, \qquad (2)$$

lo que también requiere N^2 operaciones.

La elección del método y la eficacia de éste dependen de la forma de la matriz A y, además, del tipo del ordenador.

Para varios problemas A es una *matriz enrarecida* y la mayoría de sus elementos son ceros. Las matrices de este tipo aparecen con frecuencia al realizarse la aproximación de diferencias de las ecuaciones diferenciales. Los elementos de tal matriz se calculan, corrientemente, por las fórmulas dadas, razón por la cual no han de ser conservadas en la memoria de acceso rápido del ordenador. Esta cuestión es de mucha importancia, pues el orden de las matrices de este tipo puede alcanzar hasta decenas de miles e incluso centenares de miles.

De caso particular de una matriz enrarecida sirve la *matriz de cinta*; todos sus elementos no nulos se encuentran cerca de la diagonal principal, es decir, $a_{ij} = 0$, si $|i-j| > l$, donde $l < N$. Los elementos distintos de cero se disponen en $2l+1$ diagonales, incluida la diagonal principal. Como ejemplo interviene una matriz tridiagonal

$$A = \begin{bmatrix} -c_1 & b_1 & 0 & \ldots & 0 & \\ a_2 & -c_2 & b_2 & \ldots & 0 & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & & & & 0 & a_N \ -c_N \end{bmatrix}.$$

El sistema correspondiente (1) tiene por expresión

$$\begin{aligned} -c_1u^{(1)} + b_1u^{(2)} \qquad\qquad\qquad\qquad &= f^{(1)}, \\ a_iu^{(i-1)} - c_iu^{(i)} + b_iu^{(i+1)} &= f^{(i)}, \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a_Nu^{(N-1)} - c_Nu^{(N)} \qquad\quad &= f_N, \\ i = 2, 3, \ldots, N-1. \end{aligned}$$

Esta es la ecuación de segundo orden que ha sido analizado en el cap. I, donde para su resolución se aplicó el método de factorización.

3. Ecuación operacional de primera especie. Se sabe que todo matriz $A = (a_{ij})$ $(i, j = 1, 2, \ldots, N)$ define un operador lineal A que aplica el espacio H^N en sí mismo: $Au \in H^N$ para cualquier $u \in H^N$, o bien $A: H^N \to H^N$. Viceversa e todo operador A (en cierta base $\xi_1, \ldots, \xi_N$) le corresponde una matriz $A = (a_{ij})$ de dimensión $N \times N$, donde a_{ij} es el componente del vector $A\xi_j$. Por eso, la ecuación (1) puede considerarse como una *ecuación operacional de primera especie*

$$Au = f, \qquad u, f \in H^N,$$

con el operador $A: H^N \to H^N$.

Con el fin de recalcar la equivalencia de los problemas (1) y (3), dejamos invariable la designación A tanto para la matriz como para el operador. Omitiremos el índice N en H^N y escribiremos simplemente H. El paso de (1) a la ecuación operacional resulta cómodo para la exposición de la teoría de métodos iterativos. En este caso no se emplea ninguna información concreta sobre la estructura de la matriz A.

En el espacio H introduzcamos un producto escalar $(,)$, y una norma $\| u \| = \sqrt{(u, u)}$. Supondremos que el operador A es autoconjugado y positivo: $A = A^* > 0$. Analizaremos también el espacio energético H_D con el producto escalar $(u, v)_D = (Du, v)$ y la norma $\| u \|_D = \sqrt{(Du, u)}$, donde D es un operador lineal positivo y autoconjugado: $D: H \to H$, $D = D^* > 0$.

Denotemos con (ξ_s, λ_s) $(s = 1, 2, \ldots, N)$ los vectores propios y los valores propios del operador A:

$$A\xi_s = \lambda_s \xi_s, \qquad (\xi_s, \xi_m) = \delta_{sm},$$
$$s, m = 1, 2, \ldots, N.$$

Por cuanto $A > 0$, se tiene $\lambda_s > 0$, y podemos considerar que $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$, y, por consiguiente, se verifica la desigualdad

$$\lambda_1 E \leqslant A \leqslant \lambda_N E, \qquad \lambda_1 = \min_s \lambda_s, \qquad \lambda_N = \max_s \lambda_s.$$

La razón λ_N/λ_1 lleva el nombre de *número convenido*.

En la práctica resulta más conveniente emplear la razón *inversa, es decir, el parámetro* $\xi = \lambda_1/\lambda_N$, *el cual se denominará* *medida convenida*. En lo sucesivo se mostrará que de dicho parámetro depende la convergencia de las iteraciones. El parámetro ξ para las ecuaciones en diferencias que aproximan las ecuaciones de la física matemática (por ejemplo, la ecuación de Laplace) es pequeño: $\xi \approx 10^{-2}-10^{-4}$ (el número convenido es grande).

De la fórmula $u = A^{-1}f$ se ve que

$$\| u \| \leqslant \| A^{-1} \| \| f \|, \qquad \| A^{-1} \| = 1/\lambda_1.$$

Esta desigualdad expresa la estabilidad de la solución del problema (1) respecto del segundo miembro. Si $\| A^{-1} \| = 1/\lambda_1$ es muy grande, puede suceder que el problema (3) no sea correcta, es decir, inestable con relación a los errores en la prefijación del segundo miembro, incluso a los errores de redondeo.

4. Métodos directos e iterativos. *Los métodos numéricos* de la resolución del sistema (1) se subdividen convencionalmente en dos grupos y se distinguen métodos directos y los iterativos (se tienen, por supuesto, los métodos mixtos). Los métodos *directos* permiten obtener, después de un número finito de operaciones, una solución exacta del sistema de ecuaciones, siempre que la información de entrada (el segundo miembro de la ecuación f y los elementos a_{ij} de la matriz A) viene dada con toda la exactitud y los cálculos *se realizan sin redondeo. El ejemplo más simple del método* directo es el de factorización. Por supuesto, los métodos directos dan también la solución con cierta precisión, la que depende de los errores de redondeo, es decir, del ordenador y del carácter de la estabilidad de cálculo, lo que depende, a su vez, del método mismo.

El *método iterativo* permite hallar la solución aproximada del sistema construyendo una sucesión de aproximaciones (iteraciones), a partir de cierta aproximación inicial. La *propia solución aproximada es el resultado de los cálculos* obtenido después de haberse realizado un número finito de iteraciones.

La elección de tal o cual método numérico depende de varias circunstancias: de los programas disponibles, del tipo de los cálculos y de la matriz A, etc. Expliquemos las

palabras «tipo de los cálculos». Son posibles distintos planteamientos del problema:

1) hallar la solución de un problema concreto (1);

2) hallar la solución de varias variantes del problema (1) con una misma matriz A y segundos miembros diferentes de f. Puede ocurrir que un método, no optimal para un *problema (1), resulte muy eficaz para el cálculo multivariante.*

En el cálculo multivariante se puede disminuir el número medio de operaciones para una variante, si se conservan ciertas magnitudes y no se calculan de nuevo para cada variante, lo que depende, naturalmente, del ordenador y del volumen de su memoria de acceso rápido.

De aquí está claro que la elección de un algoritmo debe depender del tipo de los cálculos, del volumen de la memoria de acceso rápido del ordenador y, naturalmente, del orden del sistema. La calidad de un algoritmo se determina por el tiempo de máquina que se exige para hallar la solución del *sistema* (1). Se *elige*, naturalmente, *un método, para el cual* el tiempo de resolución es mínimo en comparación con los otros métodos. No obstante, el tiempo de cálculo depende de varios factores, entre cuales pueden citarse el número de operaciones aritméticas y lógicas que son necesarias para obtener la solución con la exactitud prefijada, la velocidad de funcionamiento y el volumen de la memoria de *acceso rápido del ordenador, la calidad del programa.* Al estimar teóricamente la calidad de los algoritmos su comparación se realiza por el número $Q(\varepsilon)$ de operaciones aritméticas suficientes para hallar la solución del problema con la exactitud prefijada $\varepsilon > 0$.

§ 2. Métodos directos

1. Método de Gauss. Hay varias variantes de cálculo del método de Gauss basado en la idea de eliminación sucesiva. El proceso de resolución del sistema de ecuaciones algebraicas lineales $Ax = f$, o

$$\sum_{j=1}^{N} a_{ij}x_j = f_i, \quad i = 1, 2, \ldots, N, \tag{1}$$

por el método de Gauss, consta de dos etapas.

PRIMERA ETAPA (*procedimiento directo*). El sistema (1) se reduce a la forma triangular

$$x + B^{+}x = \varphi, \tag{2}$$

donde $x = (x_1, \ldots, x_N)$ y $\varphi = (\varphi_1, \ldots, \varphi_N)$ son los vectores desconocido y conocido, respectivamente, B^{+} es la matriz triangular superior.

SEGUNDA ETAPA (*procedimiento inverso*). Las incógnitas $x_N, x_{N-1}, \ldots, x_1$ se determinan por las fórmulas (2) del § 1.

Pasemos a la exposición detallada del método. El *primer paso* del método de Gauss consiste en la eliminación de la incógnita x_1 de todas las ecuaciones con excepción de la primera. Supongamos que $a_{11} \neq 0$, dividamos la primera ecuación (1) ($i = 1$) por a_{11} y escribamos el sistema (1) en la forma

$$x_1 + b_{12}x_2 + \ldots + b_{1N}x_N = \varphi_1, \quad b_{1j} = a_{1j}/a_{11},$$
$$2 \leqslant j \leqslant N, \qquad \varphi_1 = f_1/a_{11}, \tag{3}$$

$$a_{i1}x_1 + a_{i2}x_2 + \ldots + a_{iN}x_N = f_i, \quad i = 2, 3, \ldots, N. \tag{4}$$

Multipliquemos la ecuación (3) por a_{i1}, donde i es cualquiera de los números $i = 2, 3, \ldots, N$, y sustrayamos la ecuación obtenida de la i-ésima ecuación (4):

$$(a_{i2} - a_{i1}b_{12})\, x_2 + \ldots + (a_{iN} - a_{i1}b_{1N})\, x_N = f_i - a_{i1}\varphi_1,$$
$$i = 2,\ 3,\ \ldots,\ N.$$

Introduciendo las designaciones

$$a_{ij}^{(1)} = a_{ij} - a_{i1}b_{1j}, \qquad f_i^{(1)} = f_i - a_{i1}\varphi_1,$$
$$i,\ j = 2,\ 3,\ \ldots,\ N, \tag{5}$$

reescribamos el sistema obtenido de ecuaciones (que es equivalente al sistema (1)) en la forma

$$x_1 + b_{12}x_2 + \ldots + b_{1N}x_N = \varphi_1,$$
$$a_{i2}^{(1)}x_2 + \ldots + a_{iN}^{(1)}x_N = f_i^{(1)}, \qquad i = 2,\ 3,\ \ldots,\ N.$$

La primera columna de la matriz de este sistema se compone de ceros, a excepción del primer elemento para $i = 1$, $j = 1$, que es igual a uno.

El *paso segundo* consiste en la eliminación x_2 del sistema

$$\begin{aligned} a_{22}^{(1)}x_2 + \ldots + a_{2N}^{(1)}x_N &= f_2^{(1)}, \\ \ldots & \ldots \\ a_{N2}^{(1)}x_2 + \ldots + a_{NN}^{(1)}x_N &= f_N^{(1)}. \end{aligned} \qquad (6)$$

Con este objeto dividamos la primera ecuación por $a_{22}^{(1)}$:

$$x_2 + b_{23}x_3 + \ldots + b_{2N}x_N = \varphi_2,$$

$$\varphi_2 = f_2^{(1)}/a_{22}^{(1)}, \quad b_{2j} = a_{2j}^{(1)}/a_{22}^{(1)}, \quad j = 3, \ldots, N,$$

multipliquémosla después por $(-a_{i2}^{(1)})$ y sumemos con la ecuación

$$a_{i2}^{(1)}x_2 + a_{i3}^{(1)}x_3 + \ldots + a_{iN}^{(1)}x_N = f_i^{(1)}, \quad i = 3, 4, \ldots, N.$$

De resultas obtendremos un sistema

$$x_2 + b_{23}x_3 + \ldots + b_{2N}x_N = \varphi_2,$$

$$a_{i3}^{(2)}x_3 + \ldots + a_{iN}^{(2)}x_N = f_i^{(2)}, \quad i = 3, 4, \ldots, N, \qquad (7)$$

$$a_{ij}^{(2)} = a_{ij}^{(1)} - a_{i2}^{(1)}b_{2j}, \quad f_i^{(2)} = f_i^{(1)} - a_{i2}^{(1)}\varphi_2,$$

$$i = 3, 4, \ldots, N. \qquad (8)$$

Para $x_3, x_4, \ldots, x_N$ tenemos un sistema de $(N-2)$-ésimo orden análogo al sistema (6) de $(N-1)$-ésimo orden para $x_2, x_3, \ldots, x_N$.

Continuando los razonamientos, obtendremos tras el $(N-1)$-ésimo *paso* (es decir, al haber excluido $x_1, x_2, \ldots, x_{N-1}$)

$$a_{NN}^{(N-1)}x_N = f_N^{(N-1)}, \quad \text{o bien } x_N = \varphi_N, \quad \varphi_N = f_N^{(N-1)}/a_{NN}^{(N-1)}. \qquad (9)$$

Llegamos en fin al sistema (2) con la matriz triangular superior

$$\begin{aligned} x_1 + b_{12}x_2 + b_{13}x_3 + \ldots + b_{1N}x_N &= \varphi_1, \\ x_2 + b_{23}x_3 + \ldots + b_{2N}x_N &= \varphi_2, \\ \ldots & \ldots \\ x_{N-1} + b_{N-1,N}x_N &= \varphi_{N-1}, \\ x_N &= \varphi_N. \end{aligned} \qquad (10)$$

El procedimiento inverso del método de Gaus consiste en determinar todos los x_i pertenecientes al sistema (10) con la matriz triangular superior. No es difícil mostrar que el método de Gauss expuesto más arriba puede aplicarse solamente en aquel caso en que todos los menores principales son distintos de cero.

Contaremos el número de multiplicaciones y divisiones en el método de Gauss. Veamos primero el procedimiento directo. En el primer paso se requieren $Q_1 = N^2$ divisiones y multiplicaciones, el segundo paso exige $Q_2 = (N-1)^2$ operaciones, etc. En total se deben hacer N pasos del procedimiento directo realizando para ello

$$\sum_{k=1}^{N} (N-k+1)^2 = \sum_{s=1}^{N} s^2 = \frac{N(N+1)(2N+1)}{6}$$

multiplicaciones y divisiones. Es evidente que en el procedimiento inverso se deben realizar $N(N-1)/2$ multiplicaciones. De este modo, para resolver el sistema de ecuaciones (1) se necesitan $Q = N(N^2+3N-1)/3$ operaciones de multiplicación y división. Se necesitarán también aproximadamente el mismo número de las operaciones de sumación.

Demos a conocer un ejemplo de aplicación del método de Gauss. Examinemos un sistema de tres ecuaciones ($N = 3$)

$$2x_1 + 4x_2 + 3x_3 = 4, \tag{11}$$

$$3x_1 + x_2 - 2x_3 = -2, \tag{12}$$

$$4x_1 + 11x_2 + 7x_3 = 7. \tag{13}$$

PROCEDIMIENTO DIRECTO. PRIMER PASO. Dividamos la primera ecuación por $a_{11} = 2$:

$$x_1 + 2x_2 + 1{,}5x_3 = 2. \tag{14}$$

Multipliquemos (14) por -3 y sumemos con (12), a continuación multipliquemos (14) por -4 y sumemos con (13):

$$-5x_2 - 6{,}5x_3 = -8, \tag{15}$$

$$3x_2 + x_3 = 1. \tag{16}$$

Se ha obtenido el sistema de segundo orden.

SEGUNDO PASO. Dividamos (15) por -5:

$$x_2 + 1{,}3x_3 = 1{,}6. \tag{17}$$

Multipliquemos (17) por -3 y sumemos con (16):

$$-2{,}9x_3 = -5{,}8. \tag{18}$$

TERCER PASO. Dividamos (18) por $-2{,}9$:

$$x_3 = 2.$$

De resultas obtenemos un sistema

$$\begin{aligned} x_1 + 2x_2 + 1{,}5x_3 &= 2, \\ x_2 + 1{,}3x_3 &= 1{,}6, \\ x_3 &= 2 \end{aligned}$$

con la matriz triangular superior

$$\begin{bmatrix} 1 & 2 & 1{,}5 \\ 0 & 1 & 1{,}3 \\ 0 & 0 & 1 \end{bmatrix}.$$

PROCEDIMIENTO INVERSO. Del sistema hallamos sucesivamente: $x_3 = 2$, $x_2 = 1{,}6 - 1{,}3 \cdot x_3 = 1{,}6 - 1{,}3 \cdot 2 = -1$, $x_1 = 2 - 2x_2 - 1{,}5 \cdot x_3 = 1$. De este modo, queda determinada la solución del sistema (11)—(13):

$$x_1 = 1, \qquad x_2 = -1, \qquad x_3 = 2.$$

2. Método de la raíz cuadrada. Este método se emplea para los sistemas

$$Au = f \tag{19}$$

con matriz hermitiana A (simétrica, en el caso real). La matriz A se desarrolla en un producto

$$A = S^*DS, \tag{20}$$

donde S y D son matrices superior triangular y diagonal, respectivamente. La resolución de la ecuación $Au = f$ se reduce a la resolución de dos sistemas

$$S^*Dy = f, \qquad Su = y. \tag{21}$$

Con el fin de obtener el desarrollo (20), designamos $S = (s_{ij})$, $D = (d_{ii}\delta_{ij})$ y hallamos

$$(DS)_{ij} = \sum_{k=1}^{N} d_{ik}t_{kj} = d_{ii}s_{ij}, \quad (S^*DS)_{ij} = \sum_{k=1}^{N} \bar{s}_{ki}d_{kk}s_{kj},$$

puesto que $S^* = (\bar{s}_{ij})$, donde la raya significa una conjugación compleja.

De resultas obtenemos una ecuación

$$\sum_{k=1}^{N} \bar{s}_{ki}d_{kk}s_{kj} = a_{ij}. \tag{22}$$

El sistema de ecuaciones (22) se puede resolver de manera recurrente. Por cuanto S es una matriz triangular superior, entonces $s_{ki} = 0$ para $k > i$, $\bar{s}_{ik} = 0$ para $k < i$, y, por lo tanto,

$$\sum_{k=1}^{N} \bar{s}_{ki}s_{kj}d_{kk} = \sum_{k=1}^{i-1} s_{ki}s_{kj}d_{kk} + \bar{s}_{ii}s_{ij}d_{ii} + \sum_{k=i+1}^{N} s_{ki}s_{kj}d_{kk} =$$
$$= \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk} + s_{ii}s_{ij}d_{ii} = a_{ij}.$$

Para $i = j$ tenemos

$$|s_{ii}|^2 d_{ii} = a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}. \tag{23}$$

Al escoger

$$d_{ii} = \operatorname{sign}\left(a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk}\right), \tag{24}$$

hallemos

$$s_{ii} = \sqrt{\left| a_{ii} - \sum_{k=1}^{i-1} |s_{ki}|^2 d_{kk} \right|}. \tag{25}$$

Cuando $i < j$, obtenemos

$$s_{ij} = \frac{a_{ij} - \sum_{k=1}^{i-1} \bar{s}_{ki}s_{kj}d_{kk}}{\bar{s}_{ii}d_{ii}}. \tag{26}$$

Suponiendo $i = 1, 2, \ldots$, encontramos sucesivamente

$$s_{11} = \sqrt{|a_{11}|}, \quad d_{11} = \operatorname{sign} a_{11}, \quad s_{22} = \sqrt{|a_{22} - d_{11}| \, |s_{12}|^2}, \ldots$$

El determinante de la matriz es, evidentemente, igual a

$$\det A = \prod_{i=1}^{N} d_{ii} s_{ii}^2.$$

El método de la raíz cuadrada requiere aproximadamente $N^3/3$ operaciones aritméticas, es decir, cuando N es grande, el método es dos veces más rápido en comparación con el método de Gauss y ocupa dos veces menos células de la memoria. Esta circunstancia se debe a que el método emplea la información sobre la simetría de la matriz.

3. Relación del método de Gauss con el desarrollo de la matriz en factores. Sea dada una matriz regular A de dimensión $N \times N$. Representémosla en forma de un producto

$$A = BC, \qquad A = (a_{ij}), \qquad B = (b_{ij}), \qquad C = (c_{ij}) \tag{27}$$

donde B y C son las matrices triangulares de la forma

$$B = \begin{bmatrix} b_{11} & 0 & & \ldots & 0 \\ b_{21} & b_{22} & 0 & \ldots & 0 \\ b_{31} & b_{32} & b_{33} & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ b_{N1} & b_{N2} & b_{N3} & \ldots & b_{NN} \end{bmatrix},$$

$$C = \begin{bmatrix} 1 & c_{12} & c_{13} & \ldots & c_{1N} \\ 0 & 1 & c_{23} & \ldots & c_{2N} \\ 0 & 0 & 1 & \ldots & c_{3N} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & \ldots & 1 \end{bmatrix},$$

es decir, $b_{ik} = 0$ para $k > i$, $c_{ik} = 0$ para $k < i$, $c_{ii} = 1$. De (27) se infiere que

$$a_{ij} = \sum_{k=1}^{N} b_{ik} c_{kj}.$$

Transformemos esta suma de dos modos:

$$\sum_{k=1}^{N} b_{ik}c_{kj} = \sum_{k=1}^{i-1} b_{ik}c_{kj} + b_{ii}c_{ij} + \sum_{k=i+1}^{N} b_{ik}c_{kj} = \sum_{k=1}^{i-1} b_{ik}c_{kj} + b_{ii}c_{ij},$$

$$\sum_{k=1}^{N} b_{ik}c_{kj} = \sum_{k=1}^{j-1} b_{ik}c_{kj} + b_{ij}c_{jj} + \sum_{k=j+1}^{N} b_{ik}c_{kj} = \sum_{k=1}^{j-1} b_{ik}c_{kj} + b_{ij}c_{jj}.$$

De aquí encontramos

$$b_{ij} = a_{ij} - \sum_{k=1}^{j-1} b_{ik}c_{kj} \quad \text{para } i \geqslant j, \; b_{i1} = a_{i1}, \; c_{ii} = 1,$$

$$c_{ij} = \frac{1}{b_{ii}}\left[a_{ij} - \sum_{k=1}^{i-1} b_{ik}c_{kj}\right] \quad \text{para } i < j.$$

Las matrices B y C quedan determinadas.

La resolución de la ecuación $Au = BCu = f$ se reduce a la resolución sucesiva de las ecuaciones

$$B\varphi = f, \qquad Cu = \varphi.$$

La construcción de las matrices B y C, como también la búsqueda de $\varphi = B^{-1}f$ corresponden al procedimiento directo, mientras que la resolución de la ecuación

$$Cu = \varphi$$

corresponde al procedimiento inverso del método de Gauss.

§ 3. Métodos iterativos

1. Método de iteraciones para resolver el sistema de ecuaciones algebraicas lineales. Una atención especial se prestará en este capítulo a los métodos iterativos, puesto que dichos métodos son de amplio uso en la resolución de las ecuaciones en diferencias de la física matemática cuyos operadores están en correspondencia con las matrices de cinta A de orden superior.

Pasemos a la descripción general del *método de iteraciones* para un sistema de ecuaciones algebraicas lineales

$$Au = f. \qquad (1)$$

Con el fin de resolverlo se elige cierta aproximación inicial $y_0 \in H$ y se hallan sucesivamente soluciones aproximadas (iteraciones) de la ecuación (1). El valor de una *iteración* y_{k+1} se expresa en términos de las iteraciones precedentes conocidas $y_k, y_{k-1}, \ldots$. Si, al calcular y_{k+1}, se utiliza sólo una iteración precedente y_k, entonces el método iterativo se denomina *de un paso* (*de dos capas*); si, en cambio, y_{k+1} se expresa en términos de dos iteraciones, y_k e y_{k-1}, el método se llama *de dos pasos* (o *de tres capas*). En esta obra se analizarán, principalmente, los métodos de un paso. Convengamos en considerar que $A: H \to H$ es un operador lineal en un espacio de dimensión finita H con un producto escalar $(\cdot, \cdot)$.

Un papel importante lo desempeña la inscripción de los métodos iterativos en la forma unificada (canónica). Cualquier método iterativo de dos capas puede ser escrito en la siguiente forma canónica:

$$B\,\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \text{ para todos } y_0 \in H, \qquad (2)$$

donde $A: H \to H$ es el operador de la ecuación de partida (1), $B: H \to H$, operador lineal que cuenta con su inversa B^{-1}, k es el número de la iteración; $\tau_1, \tau_2, \ldots, \tau_{k+1}$ son todos los parámetros de iteración, $\tau_{k+1} > 0$. El operador B puede depender, en el caso general, del número k; para simplificar la exposición, suponemos siempre que B no depende de k.

Si $B = E$ es un operador unidad, entonces el método

$$\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \quad \text{para todos los } y_0 \in H, \qquad (3)$$

se denominará *explícito*: y_{k+1} se halla por una fórmula explícita

$$y_{k+1} = y_k - \tau_{k+1}(Ay_k - f).$$

Generalmente, cuando $B \neq E$, el método (2) se llama iterativo *implícito*: para determinar y_{k+1} hace falta resolver la ecuación

$$By_{k+1} = By_k - \tau_{k+1}(Ay_k - f) = F_k, \quad k = 0, 1, \ldots \tag{4}$$

Es natural exigir que el volumen de los cálculos para resolver el sistema $By_{k+1} = F_k$ sea inferior al volumen de los cálculos para la resolución directa del sistema $Au = f$.

La exactitud del método iterativo (2) se caracteriza por la magnitud del error $z_k = y_k - u$, es decir, por la diferencia entre la solución y_k de la ecuación (2) y la solución exacta u del sistema inicial de ecuaciones algebraicas lineales. La sustitución $y_k = z_k + u$ en (2) lleva a una ecuación homogénea para el error:

$$B\frac{z_{k+1} - z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, \ldots, \quad z_0 = y_0 - u. \tag{5}$$

Suele decirse que un método iterativo *converge* en H_D, si

$$\lim_{k\to\infty} \| z_k \|_D = 0, \text{ donde } \| z \|_D \sqrt{(Dz, z)}, \ D = D^* > 0,$$

$$D: H \to H.$$

En el caso general se prefija cierto error (relativo) $\varepsilon > 0$ con el que se debe hallar la solución aproximada y_k, los cálculos se dan por terminados cuando queda cumplida la condición

$$\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D. \tag{6}$$

Si $n = n(\varepsilon)$ es el mínimo de los números, para los cuales se verifica (6), entonces el número total de operaciones aritméticas que han de realizarse para hallar la solución aproximada de la ecuación (1) es igual a $Q_n(\varepsilon) = n(\varepsilon) q_0$, donde q_0 es el número de operaciones que se realizan para hallar una iteración, es decir, para resolver la ecuación (4). El problema consiste en minimizar $Q_n(\varepsilon)$ eligiendo de modo adecuado B y los parámetros $\{\tau_k\}$. Comencemos por analizar los métodos iterativos más simples.

2. Método de la iteración simple. Para la resolución del sistema de ecuaciones (1) puede emplearse el *método de la*

iteración simple

$$y_{k+1}^{(i)} = y_k^{(i)} - \tau \left(\sum_{j=1}^{N} a_{ij} y_k^{(j)} - f^{(i)} \right), \quad i = 1, 2, \ldots, N, \tag{7}$$

donde $\tau > 0$ es el parámetro de iteración. Escribamos (7) en la forma operacional:

$$\frac{y_{k+1} - y_k}{\tau} + A y_k = f, \quad k = 0, 1, \ldots, \text{ para cualquier } y_0 \in H. \tag{8}$$

Al comparar con (3) vemos que el método de la iteración simple se da mediante un esquema explícito de dos capas con el parámetro constante $\tau_k \equiv \tau$.

Existen también otras variantes del método de la iteración simple, por ejemplo, una que sigue:

$$y_{k+1}^{(i)} = \frac{1}{a_{ii}} \left(\sum_{j \neq i}^{1+N} a_{ij} y_k^{(j)} - f^{(i)} \right).$$

Al sustituir aquí

$$\sum_{j \neq i}^{1+N} a_{ij} y_k^{(j)} = \sum_{j=1}^{N} a_{ij} y_k^{(j)} - a_{ii} y_k^{(i)} = (A y_k)^{(i)} - (D y_k)^{(i)},$$

donde $D = (a_{ij}\delta_{ij})$ es una matriz diagonal, obtenemos

$$y_{k+1}^{(i)} = y_k^{(i)} - \frac{1}{a_{ii}} \left(\sum_{j=1}^{N} a_{ij} y_k^{(j)} - f^{(i)} \right),$$

o bien, en la forma canónica

$$D \frac{y_{k+1} - y_k}{\tau} + A y_k = f, \quad k = 0, 1, \ldots, \quad \tau = 1.$$

Aunque este esquema es formalmente implícito ($B = D \neq E$), no obstante $D = (a_{ii}\delta_{ij})$ es una matriz diagonal, por lo cual y_{k+1} se determina según las fórmulas explícitas.

3. Método de Seidel. Es de amplio uso (en particular, cuando es insuficiente la información sobre la matriz A) el

método iterativo *de Seidel* en una de las siguientes formas:

$$\sum_{j=1}^{i} a_{ij} y_{k+1}^{(j)} + \sum_{j=i+1}^{N} a_{ij} y_k^{(j)} = f^{(i)}, \quad a_{ii} \neq 0, \quad i = 1, 2, \ldots, N, \tag{9}$$

$$\sum_{j=1}^{i} a_{ij} y_k^{(j)} + \sum_{j=i+1}^{N} a_{ij} y_{k+1}^{(j)} = f^{(i)}, \quad i = 1, 2, \ldots, N. \tag{10}$$

Los componentes del vector y_{k+1} se hallan sucesivamente de ambas fórmulas. Así, por ejemplo, de (9) determinamos sucesivamente $y_{k+1}^{(1)}, y_{k+1}^{(2)}, \ldots, y_{k+1}^{(N)}$:

$$y_{k+1}^{(1)} = \frac{1}{a_{11}} \left(f^{(1)} - \sum_{j=2}^{N} a_{1j} y_k^{(j)} \right),$$

$$y_{k+1}^{(i)} = \frac{1}{a_{ii}} \left(f^{(i)} - \sum_{j=i+1}^{N} a_{ij} y_k^{(j)} - \sum_{j=1}^{i-1} a_{ij} y_{k+1}^{(j)} \right),$$
$$i = 2, \ldots, N.$$

Haciendo uso de (10) encontramos sucesivamente para $i = N, N-1, \ldots, 1$

$$y_{k+1}^{(N)} = \frac{1}{a_{NN}} \left(f^{(N)} - \sum_{i=1}^{N-1} a_{Nj} y_k^{(j)} \right),$$

$$y_{k+1}^{(i)} = \frac{1}{a_{ii}} \left(f^{(i)} - \sum_{j=1}^{i-1} a_{ij} y_k^{(j)} - \sum_{j=i+1}^{N} a_{ij} y_{k+1}^{(j)} \right),$$
$$i = N-1, \ldots, 1.$$

Escribamos este método en la forma matricial (operacional). Con este fin representemos la matriz A como una suma

$$A = A^- + D + A^+,$$

donde $D = (a_{ii}\delta_{ij})$ es una matriz diagonal de dimensión $N \times N$, $A^- = (a_{ij}^-)$ es la matriz triangular (subdiagonal) inferior con la diagonal principal llenada de ceros, $a_{ij}^- = 0$ para $j \geqslant i$, $a_{ij}^- = a_{ij}$ para $j < i$, $A^+ = (a_{ij}^+)$ es la matriz triangular (sobrediagonal) superior con la diagonal principal

llenada de ceros, $a_{ij}^{+}=0$ para $j\leqslant i$, $a_{ij}^{+}=a_{ij}$ para $j>i$. De la definición de A^{-}, D, A^{+} se desprende que

$$Dy^{(i)}=a_{ii}y^{(i)},\quad A^{-}y^{(i)}=\sum_{j=1}^{i-1}a_{ij}y^{(j)},$$

$$A^{+}y^{(i)}=\sum_{j=i+1}^{N}a_{ij}y^{(i)},\quad (A^{+}+D)\,y^{(i)}=\sum_{j=i}^{N}a_{ij}y^{(j)}.$$

Por esto, la ecuación (10) puede anotarse en la forma

$$((A^{+}+D)\,y_{k+1})^{(i)}+(A^{-}y_k)^{(i)}=f^{(i)},\quad i=1,\ 2,\ \ldots,\ N,$$

o bien, en la forma vectorial,

$$(A^{+}+D)\,y_{k+1}+A^{-}y_k=f.$$

Realizadas unas transformaciones evidentes

$$(A^{+}+D)\,y_{k+1}+A^{-}y_k=(A^{+}+D)\,(y_{k+1}-y_k)+$$
$$+(A^{-}+(A^{+}+D))\,y_k=(A^{+}+D)\,(y_{k+1}-y_k)+Ay_k$$

escribamos el método de Seidel (10) en la forma canónica:

$$(D+A^{+})\,(y_{k+1}-y_k)+Ay_k=f,\quad k=0,\ 1,\ 2,\ \ldots \tag{11}$$

Comparando con (2) vemos que el método de Seidel (10) corresponde a

$$B=D+A^{+},\qquad \tau\equiv 1,$$

es decir, el esquema (11) es implícito. Sin embargo, por cuanto $B=D+A^{+}$ es una matriz triangular, entonces la iteración y_{k+1} se determina por las fórmulas explícitas. Análogamente se escribe la otra variante del método de Seidel:

$$(D+A^{-})\,(y_{k+1}-y_k)+Ay_k=f,\qquad k=0,\ 1,\ \ldots, \tag{12}$$

cuando $B=D+A^{-}$ es una matriz triangular inferior. A continuación (en el p. 5) se mostrará que el método de Seidel converge, si A es una matriz simétrica definida positiva.

4. Método de relajación superior. Con el objeto de acelerar un proceso iterativo, se puede reducir el método de Seidel al de relajación superior, introduciendo el parámetro

de iteración ω de modo tal que se verifique

$$(D+\omega A^-)\frac{y_{k+1}-y_k}{\omega}+Ay_k=f,$$
$$k=0,\ 1,\ \ldots,\ \text{para cualquier } y_0\in H. \quad (13)$$

Al comparar con (2) vemos que

$$B=D+\omega A^-, \qquad \tau=\omega.$$

Transformemos la ecuación (13) a la forma de cálculo. Teniendo presente que

$$(D+\omega A^-)\frac{y_{k+1}-y_k}{\omega}+Ay_k=\left(A^-+\frac{1}{\omega}D\right)y_{k+1}+$$
$$+\left(A-A^--\frac{D}{\omega}\right)y_k=\left(A^-+\frac{1}{\omega}D\right)y_{k+1}+$$
$$+\left(A^++\left(1-\frac{1}{\omega}\right)D\right)y_k,$$

tenemos

$$\left(A^-+\frac{1}{\omega}D\right)y_{k+1}+\left(A^++\left(1-\frac{1}{\omega}\right)D\right)y_k=f.$$

De aquí encontramos

$$y_{k+1}^{(i)}=y_k^{(i)}+\frac{\omega}{a_{ii}}\left[f^{(i)}-\sum_{j=1}^{i-1}a_{ij}y_{k+1}^{(j)}-\sum_{j=i}^{N}a_{ij}y_k^{(j)}\right],$$
$$i=1,\ 2,\ \ldots,\ N.$$

Cuando $\omega=1$, obtenemos la fórmula del método de Seidel.

La velocidad de convergencia del método de relajación superior depende del parámetro ω. En el p. 5 se mostrará que para la convergencia del método se ha de exigir que $0<\omega<$ <2.

5. Convergencia de los métodos iterativos estacionarios. El método de Seidel y el de relajación superior sirven de ejemplo de los esquemas implícitos de la forma

$$B\frac{y_{k+1}-y_k}{\tau}+Ay_k=f,\quad k=0,\ 1,\ \ldots,$$
$$\text{para cualquier } y_0\in H, \quad (14)$$

con el operador no autoconjugado B que tiene su inverso B^{-1}. El método (14) lleva el nombre de iterativo *estacionario*, puesto que B y τ no dependen del número de iteración. Para que exista el operador inverso B^{-1}, es suficiente exigir que el operador B sea positivo. Sea $B = D + \omega A^-$. Por cuanto $A = A^* > 0$, entonces $(A^- y, y) = (A^+ y, y)$, $(A^+)^* = A^-$, y, por consiguiente, $(Ay, y) = (Dy, y) + 2(A^- y, y)$, es decir,

$$(A^- y, y) = \frac{1}{2}((A - D)y, y).$$

Sustituyendo esta expresión en la fórmula $(By, y) = (Dy, y) + \omega(A^- y, y)$, hallamos

$$(By, y) = \left(1 - \frac{1}{2}\omega\right)(Dy, y) + \omega(Ay, y) > 0,$$

siempre que $0 < \omega < 2$.

Para el error $z_k = y_k - u$ obtenemos una ecuación homogénea

$$B\frac{z_{k+1} - z_k}{\tau} + Az_k = 0, \quad k = 0, 1, 2, \ldots, z_0 = y_0 - u. \quad (15)$$

TEOREMA 1. *Sea A un operador autoconjugado y positivo y suponemos cumplida la condición*

$$B > \frac{\tau}{2}A. \quad (16)$$

En este caso el método de iteraciones (14) *converge en H_A, es decir,*

$$\| z_k \|_A = \| y_k - u \|_A \to 0 \quad \textit{cuando} \quad k \to \infty.$$

DEMOSTRACIÓN. Nos hará falta una identidad energética

$$2\tau\left(\left(B - \frac{\tau}{2}A\right)\frac{z_{k+1} - z_k}{\tau}, \frac{z_{k+1} - z_k}{\tau}\right) + \| z_{k+1} \|_A^2 = \| z_k \|_A^2, \quad (17)$$

donde $\| z \|_A^2 = (Az, z)$. Transformemos primero la ecuación (15) a la forma

$$\left(B - \frac{\tau}{2}A\right)\frac{z_{k+1} - z_k}{\tau} + \frac{1}{2}A(z_k + z_{k+1}) = 0, \quad (18)$$

sustituyendo con este fin $z_k = \frac{1}{2}(z_{k+1}+z_k) - \frac{\tau}{2}\frac{(z_{k+1}-z_k)}{\tau}$.

Al multiplicar (18) escalarmente por $2\tau\left(\frac{z_{k+1}-z_k}{\tau}\right) = 2(z_{k+1}-z_k)$ y teniendo presente que $(Az_{k+1}, z_k) = (z_{k+1}, Az_k)$, puesto que $A = A^*$ y $(A(z_k+z_{k+1}), z_{k+1}-z_k) = (Az_{k+1}, z_{k+1}) - (Az_k, z_k) + (Az_k, z_{k+1}) - (Az_{k+1}, z_k) = (Az_{k+1}, z_{k+1}) - (Az_k, z_k)$, obtenemos (17).

Supongamos cumplida la condición $B > \tau A/2$. Entonces, el primer sumando en el miembro izquierdo de la identidad (17) es no negativo y $\| z_{k+1}\|_A^2 \leqslant \| z_k \|_A^2$. De aquí se deduce que $0 \leqslant \| z_{k+1}\|_A \leqslant \| z_k\|_A \leqslant \ldots \leqslant \| z_0 \|_A$, es decir, la sucesión $\{\| z_k \|_A\}$ no es creciente y está acotada inferiormente por cero. Por ello, en virtud del teorema de Weierstrass, $\{\| z_k\|_A\}$ converge para $k\to\infty$. Demostremos que $\lim\limits_{k\to\infty} \| z_k \|_A = 0$.

El operador $P = B - \frac{\tau}{2}A$ es positivo, y $P_0 = B_0 - \frac{\tau}{2}A = \frac{1}{2}(P+P^*)$, definido positivo, es decir, existe tal número $\delta > 0$ (véase el cap. I, § 4), que

$$(Py, y) = (P_0 y, y) \geqslant \delta \| y \|^2 \text{ para cualquier } y \in H.$$

Por eso, de la identidad (17) obtenemos una desigualdad

$$\frac{2\delta}{\tau}\| z_{k+1} - z_k \|^2 + \| z_{k+1}\|_A^2 \leqslant \| z_k \|_A^2. \qquad (*)$$

Dado que $\{\| z_k \|_A\}$ es convergente, de aquí se infiere que existe

$$\lim_{k\to\infty} \| z_{k+1} - z_k \| = 0. \qquad (19)$$

Luego, de la ecuación (15) encontramos

$$Az_k = -\frac{1}{\tau} B(z_{k+1}-z_k), \quad z_k = -\frac{1}{\tau} A^{-1}B(z_{k+1}-z_k),$$

$$(Az_k, z_k) = \frac{1}{\tau^2}(A^{-1}B(z_{k+1}-z_k), B(z_{k+1}-z_k)),$$

$$\| z_k \|_A^2 \leqslant \frac{1}{\tau^2} \| A^{-1} \| \| B \|^2 \| z_{k+1} - z_k \|^2 \qquad (**)$$

De aquí precisamente concluimos que $\lim_{k\to\infty} \| z_k \|_A = 0$.

OBSERVACIÓN. De las desigualdades (*) y (**) proviene que el método de iteraciones (14) converge en las condiciones (16) con la velocidad de una progresión geométrica, $\| z_{k+1} \|_A^2 \leqslant \rho^2 \| z_k \|_A^2$, donde $\rho^2 = 1 - \frac{2\delta\tau}{\|A^{-1}\| \, \|B\|^2} < 1$.

Apliquemos el teorema 1 para demostrar la convergencia de los métodos iterativos estudiados en los pp. 2—4.

MÉTODO DE LA ITERACIÓN SIMPLE, $B = E$. Al tomar en consideración que $E \geqslant \frac{1}{\|A\|} A$, tenemos

$$B - \frac{\tau}{2} A = B - \frac{\tau}{2} A \geqslant \left(\frac{1}{\| A \|} - \frac{\tau}{2} \right) A > 0$$

para $\frac{1}{\| A \|} - \frac{\tau}{2} > 0$. El método de la iteración simple converge para todos los valores de τ, que satisfacen la desigualdad $\tau < 2/\| A \|$.

MÉTODO DE SEIDEL, $B = D + A^-$, $\tau = 1$. En este caso

$$B - \frac{1}{2} A = D + A^- - \frac{1}{2} (A^- + A^+ + D) = \frac{D}{2} + \frac{1}{2} (A^- - A^+),$$

$$\left(\left(B - \frac{1}{2} A \right) y, \, y \right) = \frac{1}{2} (Dy, \, y) + \frac{1}{2} ((A^+ - A^-) y, \, y) = \\ = \frac{1}{2} (Dy, \, y) > 0,$$

siempre que $D > 0$.

OBSERVACIÓN. La desigualdad $D > 0$ proviene de la condición $A > 0$. En efecto, supongamos que $A > 0$ y $\xi = (\xi^1, 0, \ldots, 0)$; entonces $(A\xi, \xi) = (D\xi, \xi) = a_{11} (\xi^1)^2 > 0$, es decir, $a_{11} > 0$. De un modo análogo nos convencemos de que $a_{ii} > 0$, y, por consiguiente, $D > 0$. Así pues, el método de Seidel es siempre convergente, si A es un operador autoconjugado positivo.

Para estimar la velocidad de convergencia se deben estipular las suposiciones más fuertes. Citemos el siguiente

TEOREMA 2. *El método de Seidel converge con la velocidad de una progresión geométrica de razón* $q < 1$, si $A = (a_{ij}) = A^* > 0$, y

$$\sum_{j \neq i}^{1 \div N} |a_{ij}| \leqslant q \, |a_{ii}|, \quad i = 1, \, 2, \, \ldots, \, N, \quad q < 1. \qquad (20)$$

En efecto, para el error $z_k = y_k - u$ tenemos

$$a_{ii} z_{k+1}^{(i)} = -\sum_{j<i} a_{ij} z_{k+1}^{(j)} - \sum_{j>i} a_{ij} z_k^{(j)},$$

$$|a_{ii}|\, |z_{k+1}^{(i)}| \leqslant \sum_{j<i} |a_{ij}|\, |z_{k+1}^{(i)}| + \sum_{j>i} |a_{ij}|\, |z_k^{(j)}|.$$

Supongamos que el máx $|z_{k+1}^{(i)}|$ se alcanza para cierto $i = i_0$, de modo que

$$\| z_{k+1} \|_C = |z_{k+1}^{(i_0)}|, \quad |a_{i_0 i_0}| \cdot \| z_{k+1} \|_C \leqslant \sum_{j<i_0} |a_{i_0 j}| \cdot \| z_{k+1} \|_C + \sum_{j>i_0} |a_{i_0 j}|\, \| z_k \|_C,$$

$$\| z_{k+1} \|_C \leqslant \Big[\sum_{j>i_0} |a_{i_0 j}| / \Big(|a_{i_0 i_0}| - \sum_{j<i_0} |a_{i_0 j}| \Big) \Big] \| z_k \|_C.$$

En virtud de la condición (20) tenemos

$$\sum_{j>i_0} |a_{i_0 j}| \leqslant q\, |a_{i_0 i_0}| - \sum_{j<i_0} |a_{i_0 j}| < q \Big(|a_{i_0 i}| - \sum_{j<i_0} |a_{i_0 j}| \Big),$$

y, por consiguiente,

$$\| z_{k+1} \|_C \leqslant q\, \| z_k \|_C \leqslant q^{k+1} \| z_0 \|_C,$$

lo que se trataba de demostrar.

La condición (20) significa que $A = (a_{ij})$ es una matriz con preponderancia diagonal.

MÉTODO DE RELAJACIÓN SUPERIOR. $B = D + \omega A^-$, $\tau = \omega$. Hallemos la diferencia

$$B - \frac{\tau}{2} A = D + \omega A^- - \frac{\omega}{2} (A^- + A^+ + D) = \left(1 - \frac{\omega}{2}\right) D + \frac{\omega}{2} (A^- - A^+)$$

y calculemos

$$\left(\left(B - \frac{\tau}{2} A\right) y,\ y \right) = \left(1 - \frac{\omega}{2}\right) (Dy,\ y) > 0 \text{ para } 0 < \omega < 2.$$

De este modo el método de relajación superior converge para cualesquiera valores de $\omega \in (0, 2)$, si $A = A^* > 0$.

6. Velocidad de convergencia del método implícito de iteración simple. El propio hecho de convergencia de las iteraciones no es bastante para poder juzgar sobre la aplica-

bilidad en la práctica de tal o cual método iterativo. Se necesita la información sobre la velocidad de la convergencia del método, es decir, de hecho, sobre el número de iteraciones $n = n_0(\varepsilon)$ que sean suficientes para la resolución del problema con una exactitud prefijada $\varepsilon > 0$. El número de iteraciones $n_0(\varepsilon)$ depende del parámetro τ, el que debe precisamente escogerse a partir de la condición del número mínimo de iteraciones $n = n(\varepsilon)$, con el cual se cumple la condición $\| y_n - u \|_D \leqslant \varepsilon \| y_0 - u \|_D$, donde D es un operador, $D = D^* > 0$.

Analizaremos aquí un esquema estacionario implícito (esquema implícito de la iteración simple)

$$B \frac{y_{k+1} - y_k}{\tau} + Ay_k = f, \quad k = 0, 1, \ldots, \quad \text{para cualesquiera } y_0 \in H, \tag{21}$$

donde A y B son los operadores autoconjugados positivos.

Los métodos de Seidel y de relajación superior no pertenecen a esta familia de esquemas, puesto que para ellos el operador B no es autoconjugado. Para la corrección

$$w_k = B^{-1} r_k, \qquad r_k = Ay_k - f$$

se verifica (al igual que para el error $z_k = y_k - u$) una ecuación homogénea

$$B \frac{w_{k+1} - w_k}{\tau} + Aw_k = 0, \quad k = 0, 1, \ldots, \quad w_0 = B^{-1}(Ay_0 - f) \tag{22}$$

donde $r_k = Ay_k - f$ es un defecto, $w_k = B^{-1} r_k$ es la corrección. Efectivamente, de (21) encontramos

$$y_{k+1} = y_k - \tau B^{-1}(Ay_k - f) = y_k - \tau w_k,$$

$$Ay_{k+1} - f = Ay_k - f - \tau A w_k, \qquad r_{k+1} = r_k - \tau A w_k.$$

Por cuanto $r_k = B(B^{-1} r_k) = Bw_k$, de aquí proviene (22).

Supondremos cumplidas las desigualdades operacionales

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \quad \gamma_1 > 0, \quad \gamma_2 \geqslant \gamma_1 > 0, \tag{23}$$

o bien

$$\gamma_1 (Bx, x) \leqslant (Ax, x) \leqslant \gamma_2 (Bx, x) \quad \text{para cualesquiera } x \in H, \tag{24}$$

donde las constantes γ_1, γ_2 son conocidas.

TEOREMA 3. *Supongamos cumplidas las condiciones* (23), (24). *En este caso el número mínimo de iteraciones según el método* (21) *se alcanza para*

$$\tau = \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \tag{25}$$

Además, se verifica la desigualdad

$$\| Ay_n - f \|_{B^{-1}} \leqslant \rho_0^n \| Ay_0 - f \|_{B^{-1}}, \quad n = 1, 2, \ldots, \tag{26}$$

$$\rho_0 = (1 - \xi)/(1 + \xi), \quad \xi = \gamma_1/\gamma_2. \tag{27}$$

DEMOSTRACIÓN. Con el fin de resolver el problema (22) hagamos uso de la siguiente estimación (la demostración de la estimación se aduce en el cap. V)

$$\| w_n \|_B \leqslant \rho^n \| w_0 \|_B \quad \text{para} \quad \tau \leqslant \tau_0, \tag{28}$$

donde $\rho = 1 - \tau\gamma_1$. El valor mínimo de ρ (para el cual el número de iteraciones es mínimo) se alcanza si $\tau = \tau_0$: $\rho \geqslant \rho_0 = 1 - \tau_0\gamma_1 = (1 - \xi)/(1 + \xi)$. Nos queda tomar en consideración que $\| w_n \|_B = \| B^{-1} r_n \|_B = \| r_n \|_{B^{-1}}$. El teorema está demostrado.

Exigiendo que sea $\rho_0^n \leqslant \varepsilon$, o bien $(1/\rho_0)^n \geqslant 1/\varepsilon$, obtendremos la estimación para el número de iteraciones:

$$n \geqslant \ln (1/\varepsilon)/\ln (1/\rho_0). \tag{29}$$

OBSERVACIÓN. Una función $\varphi(\xi) = \ln (1 + \xi)/(1 - \xi) - 2\xi$ es positiva para cualesquiera $0 < \xi < 1$, puesto que $\varphi'(\xi) = 2\xi^2/(1 - \xi^2) > 0$, $\varphi(0) = 0$; por esto, $1/\ln (1/\rho_0) < 1/(2\xi)$ y la condición (29) queda cumplida, siempre que

$$n \geqslant n_0(\varepsilon) = (1/(2\xi)) \ln 1/\varepsilon, \qquad \xi = \gamma_1/\gamma_2 \tag{30}$$

($n_0(\varepsilon)$ no es, en el caso general, entero). La condición (30) resulta más cómoda para las estimaciones. La estimación $\rho_0^n \leqslant \varepsilon$ es, evidentemente, verídica, si $n_0(\varepsilon) \leqslant n < n_0(\varepsilon) + 1$. Por esta razón, a título de n es suficiente tomar la parte entera del número $n_0(\varepsilon) + 1$.

7. Problema modelo. La comparación de los diferentes métodos iterativos se realizara a base del siguiente problema modelo

$$\frac{v_{i-1} - 2v_i + v_{i+1}}{h^2} = -\tilde{f}_i, \quad i = 1, 2, \ldots, N-1,$$

$$v_0 = \mu_1, \quad v_N = \mu_2, \quad h = \frac{1}{N}, \tag{31}$$

el cual es un esquema de diferencias para el problema de contorno

$$\frac{d^2u}{dx^2} = -\tilde{f}(x), \quad 0 < x < 1, \quad u(0) = \mu_1, \quad u(1) = \mu_2.$$

Escribamos el sistema de ecuaciones primeramente en la forma matricial:

$$Av = f, \tag{32}$$

donde $v = (v^{(1)}, v^{(2)}, \ldots, v^{(N-1)})$ es un vector de dimensión $N - 1$, y A es una matriz tridiagonal de dimensión $(N-1) \times \times (N-1)$:

$$A = -\frac{1}{h^2}\begin{bmatrix} -2 & 1 & 0 \ldots & & & 0 \\ 1 & -2 & 1 \ldots & & & 0 \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 0 & & \ldots 1 & -2 & & 1 \\ 0 & & \ldots 0 & 1 & & -2 \end{bmatrix}.$$

El segundo miembro de la ecuación (32) cuenta con los componentes $f_i = \tilde{f}_i$ para $i = 2, 3, \ldots, N-2$, $\tilde{f}_1 = \tilde{f}_1 + + \mu_1/h^2$, $f_{N-1} = \tilde{f}_{N-1} + \mu_2/h^2$. A la matriz A le corresponde *el operador A que actúa en el espacio $H = \Omega$ de* funciones reticulares definidas en los nodos interiores de la red $\omega_h = = \{x_i = ih,\ 0 < i < N\}$. Sea $\Lambda v = v_{\bar{x}x}$, $\overset{\circ}{v}$ una función reticular que está definida sobre la red $\bar{\omega}_h = \{x_i = ih,$ $0 \leqslant i \leqslant N\}$ y que se reduce a cero en la frontera cuando $i = 0, N$. En este caso podemos escribir

$$Av = -\Lambda\overset{\circ}{v}, \qquad v \in \Omega = H, \qquad v \in \overset{\circ}{\Omega}.$$

Introduzcamos en $H = \Omega$ (como lo hacemos habitualmente) un producto escalar

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h$$

y hagamos uso de las fórmulas (17), (56) del § 4, cap. I, en virtud de las cuales

$$(Av, w) = (v, Aw), \quad \text{es decir,} \quad A = A^*,$$

$$(Av, v) \geqslant \delta \| v \|^2, \quad \delta = \frac{4}{h^2}\,\text{sen}^2\frac{\pi h}{2}, \quad A \geqslant \delta E.$$

Ahora tenemos

$$\| A \| = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}.$$

Estimemos el número de iteraciones para el esquema explícito de una iteración simple en el caso del problema modelo. Se tiene $B = E$, $\delta E \leqslant A \leqslant \Delta E$, es decir,

$$\gamma_1 = \delta, \quad \gamma_2 = \Delta, \quad \xi = \frac{\gamma_1}{\gamma_2} = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}.$$

Para el número de iteraciones tenemos

$$n(\varepsilon) \geqslant n_0(\varepsilon) = \frac{\ln 1/\varepsilon}{2\xi} \approx \frac{2}{10h^2} \ln \frac{1}{\varepsilon}.$$

Prefijemos $\varepsilon = \frac{1}{2} \cdot 10^{-4} \approx e^{-10}$, entonces $n_0(\varepsilon) \approx \frac{2}{h^2} = 2N^2$. En particular, el número de iteraciones:

$$n_0(\varepsilon) \approx 200 \text{ para } N = 10$$

$$n_0(\varepsilon) \approx 20\,000 \text{ para } N = 100.$$

El método de iteración simple depende fuertemente del número de ecuaciones N ($n_0(\varepsilon) \approx N^2$). Abajo se exponen los métodos (véanse los §§ 4, 5), para los cuales la dependencia citada (n en función de N) será más débil ($n_0(\varepsilon) \approx N$ y $n_0(\varepsilon) \approx \sqrt{N}$).

El problema (31) es un problema tipo, puesto que una ecuación en diferencias análoga simula la ecuación de Laplace en diferencias para los casos bidimensional y tridimensional, y el número de iteraciones no depende prácticamente del número de mediciones (depende sólo de h).

8. Esquema de tres capas. Si y_{k+1} se calcula mediante dos iteraciones precedentes y_k e y_{k-1}, entonces el método iterativo se denomina *de dos pasos* (o *de tres capas*). Demos un ejemplo del esquema iterativo de tres capas. El esquema explícito de tres capas con parámetros constantes se anota, corrientemente, en la forma

$$y_{k+1} = (1 + \alpha)(E - \tau_0 A) y_k - \alpha y_{k-1} + (1 + \alpha) \tau_0 f,$$
$$k = 1, 2, \ldots \qquad (33)$$

La primera iteración se calcula según el método explícito de iteración simple:

$$y_1 = (E - \tau_0 A)\, y_0 + \tau_0 f \quad \text{para cualesquiera} \quad y_0 \in H, \tag{34}$$

donde

$$\tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \quad \alpha = \rho_1^2, \quad \rho_1 = \frac{1 - \sqrt{\xi}}{1 + \sqrt{\xi}}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \tag{35}$$

$\gamma_1, \gamma_2 > 0$ son las fronteras del espectro del operador $A = A^*$: $\gamma_1 E \leqslant A \leqslant \gamma_2 E$.

Se puede mostrar que para el método (33), (34) el número de iteraciones se halla de la condición

$$q_n = \rho_1^n \left(1 + \frac{1-\rho_1^2}{1+\rho_1^2}\, n\right) \leqslant \varepsilon.$$

De aquí se ve que

$$n_0(\varepsilon) \approx \frac{c_0}{2\sqrt{\xi}} \ln \frac{1}{\varepsilon}, \quad 1 < c_0 < 2. \tag{36}$$

Para el problema modelo $\sqrt{\xi} \approx \pi h/2$ y

$$n_0(\varepsilon) = \frac{c_0}{\pi h} \ln \frac{1}{\varepsilon} \approx c_0 \frac{0{,}32}{h} \ln \frac{1}{\varepsilon} \approx c_0 \cdot 3{,}2N \quad \text{para } \varepsilon \approx e^{-10}.$$

El número de iteraciones:

$$n_0(\varepsilon) \approx 32 \div 60 \quad \text{para } N = 10,$$
$$n_0(\varepsilon) \approx 320 \div 620 \quad \text{para } N = 100,$$

es decir, considerablemente menos que para la iteración simple.

El esquema implícito de tres capas tiene por expresión

$$By_{k+1} = (1+\alpha)(B - \tau_0 A)\, y_k - \alpha B y_{k-1} + (1+\alpha)\,\tau_0 f,$$
$$k = 1, 2, \ldots,$$
$$By_1 = By_0 - \tau_0 A y_0 + \tau_0 f \quad \text{para cualesquiera} \quad y_0 \in H.$$

Si $B = B^* > 0$ y se cumplen las desigualdades (23), (24) mientras que α, τ_0 se calculan según las fórmulas (35), entonces la estimación (36) para el número de iteraciones queda justa en este caso también.

§ 4. Esquema iterativo de dos capas con parámetros de Chébishev

1. Planteamiento del problema. Sea dada una ecuación

$$Au = f, \qquad A\colon H \to H. \tag{1}$$

Veamos un esquema iterativo con parámetros variables $\{\tau_k\}$:

$$B\,\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, 2, \ldots,$$
$$\text{para cualquier } y_0 \in H. \tag{2}$$

La ecuación homogénea

$$B\,\frac{z_{k+1}-z_k}{\tau_{k+1}} + Az_k = 0, \quad k = 0, 1, 2, \ldots, \quad z_0 = y_0 - u, \tag{3}$$

es satisfecha no sólo por el error $z_k = y_k - u$, sino también por la corrección $w_k = B^{-1}(Ay_k - f)$ $(k = 0, 1, \ldots)$ con la condición inicial $w_0 = B^{-1}(Ay_0 - f)$. La condición para que terminen las iteraciones tiene por expresión

$$\| z_n \|_D \leqslant \varepsilon \| z_0 \|_D, \quad \text{o bien} \quad \| w_n \|_D \leqslant \varepsilon \| w_0 \|_D. \tag{4}$$

De (3) se ve que

$$z_{k+1} = S_{k+1} z_k, \qquad S_{k+1} = E - \tau_{k+1} B^{-1} A, \tag{5}$$

donde S_{k+1} es el operador de paso de la capa k a la capa $k+1$. Eliminando $z_k, z_{k-1}, \ldots, z_1$, encontramos para $k = n - 1$:

$$z_n = T_n z_0, \qquad T_n = S_n S_{n-1} \ldots S_2 S_1,$$

donde T_n es el *operador de resolución* del esquema (3). De aquí proviene que

$$\| z_n \|_D \leqslant q_n \| z_0 \|_D, \qquad q_n = \| T_n \|_D. \tag{6}$$

La condición para que terminen las iteraciones queda cumplida, si

$$q_n = \| T_n \|_D \leqslant \varepsilon. \tag{7}$$

Para estimar el número de iteraciones $n = n(\varepsilon)$ se debe obtener la desigualdad (7).

Estudiemos el esquema explícito (2)

$$\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f, \quad k=0,\ 1,\ 2,\ \ldots, \tag{8}$$

con la particularidad de que consideramos prefijado cualquier $y_0 \in H$, y elijamos los parámetros $\tau_1, \tau_2, \ldots, \tau_n$ partiendo de la condición de mín n (ε). Se supone, además, que

$$A = A^* > 0, \qquad \gamma_1 E \leqslant A \leqslant \gamma_2 E, \qquad \gamma_1 > 0.$$

Para el residuo $r_k = Ay_k - f$ se verifica una ecuación homogénea

$$\frac{r_{k+1}-r_k}{\tau_{k+1}}+Ar_k=0, \quad k=0,\ 1,\ 2,\ \ldots,\ r_0=Ay_0-f,$$

o bien

$$r_{k+1} = S_{k+1}r_k, \qquad S_{k+1} = E - \tau_{k+1}A.$$

De aquí hallemos

$$r_n = T_n r_0, \qquad T_n = S_1 S_2 \ldots S_n.$$

El operador de resolución T_n es un polinomio de grado n respecto de A:

$$T_n = P_n(A) = (E - \tau_1 A)(E - \tau_2 A) \ldots (E - \tau_n A)$$

con los coeficientes que sólo dependen de $\tau_1, \tau_2, \ldots, \tau_n$.

Para determinar $\tau_1, \tau_2, \ldots, \tau_n$ obtenemos la estimación

$$\| r_n \| \leqslant \| P_n(A) \| \, \| r_0 \|.$$

Es menester hallar tales $\tau_1, \tau_2, \ldots, \tau_n$, para los cuales $\| P_n(A) \|$ es mínima y, después, estimar dicha norma a través de las constantes γ_1 y γ_2. Demos aquí sin demostración la solución de este problema.

Designemos con $\mathfrak{M}_n = \left\{ -\cos\frac{2i-1}{2n}\pi,\ i=1,\ 2,\ \ldots,\ n \right\}$ el conjunto de ceros del *polinomio de Chébishev* $T_n(x) = \cos(n \arccos x)$ en el segmento $-1 \leqslant x \leqslant 1$, y con $\{\mu_k\}$, una sucesión cualquiera de estos ceros, $\mu_k \equiv \mathfrak{M}_n$. El número mínimo de iteraciones se alcanza para los valores de los

parámetros

$$\tau_k = \frac{\tau_0}{1+\rho_0\mu_k}, \qquad k=1,\ 2,\ \ldots,\ n,$$
$$\tau_0 = \frac{2}{\gamma_1+\gamma_2}, \qquad \rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2}. \tag{9}$$

En este caso es válida la estimación

$$\|Ay_k - f\| \leqslant q_n \|Ay_0 - f\|, \qquad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}},$$
$$\rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}. \tag{10}$$

El esquema (8) con parámetros iterativos (9) lleva el nombre de *esquema iterativo de Chébishev*.

El requisito $q_n \leqslant \varepsilon$, ó $2\rho_1^n \leqslant \varepsilon\,(1+\rho_1^{2n})$ queda cumplido, si $\rho_1^n \leqslant \varepsilon/2$, o bien

$$n(\varepsilon) \geqslant \ln\frac{2}{\varepsilon}\Big/\ln\frac{1}{\rho_1}. \tag{11}$$

Al observar (compárese con el § 3, p. 6) que $\ln\frac{1}{\rho_1} = \ln\frac{1+\sqrt{\xi}}{1-\sqrt{\xi}} > 2\sqrt{\xi}$, sustituyamos (11) por el requisito más fuerte:

$$n(\varepsilon) > n_0(\varepsilon) = \frac{1}{1\sqrt{\xi}}\ln\frac{2}{\varepsilon}, \tag{12}$$

que sea más cómodo para la comprobación. De (12) se deduce, evidentemente, (10), y $q_n \leqslant \varepsilon$.

Comparemos, según el número de iteraciones, el esquema (8) con la totalidad indicada de parámetros y el método de iteración simple, recurriendo con este fin al ejemplo del problema modelo citado en el § 3. En este caso $\xi \approx \pi^2h^2/4$, $\sqrt{\xi} = \pi h/2$.

Para el método de iteración simple

$$n_0^{(1)}(\varepsilon) \approx 2/h^2 \quad \text{para} \quad \varepsilon = 10^{-4}.$$

Para el esquema de Chébishev

$$n_0^{(2)}(\varepsilon) = 3{,}4/h \quad \text{para} \quad \varepsilon = 10^{-4}.$$

De aquí se ve que

$$n_0^{(2)} \approx 34, \quad n_0^{(1)} \approx 200 \quad \text{para } N = 10 \ (h = 1/10).$$
$$n_0^{(2)} \approx 340, \ n_0^{(1)} \approx 20\,000 \ \text{para } N = 100 \ (h = 1/100).$$

2. Argumentación de la elección optimal de los parámetros. Demostremos la estimación (10) en el caso de parámetros iterativos (9). Nos hace falta hallar el $\min\limits_{\{\tau_k\}} \| P_n(A) \|$.

El polinomio

$$P_n(A) = \prod_{k=1}^{n} (E - \tau_k A) =$$
$$= c_0 + c_1 A + \ldots + c_k A^k + \ldots + c_n A^n,$$
$$c_0 = 1, \qquad P_n(0) = 1$$

es un operador autoconjugado. Sean ξ_s, λ_s ($s = 1, 2, \ldots, N$) funciones propias y valores propios, respectivamente, del operador A:

$$A\xi_s = \lambda_s \xi_s, \quad (\xi_s, \xi_m) = \delta_{sm}, \quad s, m = 1, 2, \ldots, N.$$

El operador A^k tiene las mismas funciones propias y los mismos valores propios λ_s^k:

$$A^k \xi_s = \lambda_s^k \xi_s. \tag{13}$$

Multiplicando (13) por c_k y sumando según $k = 0, 1, \ldots, n$ ($c_0 = 1$), obtendremos

$$P_n(A)\,\xi_s = \sum_{k=0}^{n} c_k A^k \xi_s = \sum_{k=0}^{n} c_k \lambda_s^k \xi_s = P_n(\lambda_s)\,\xi_s.$$

Al cotejar esto con $P_n(A)\,\xi_s = \lambda_s (P_n(A))\,\xi_s$, vemos que

$$\lambda (P_n(A)) = P_n(\lambda(A)).$$

Los valores propios del polinomio operacional $P_n(A)$ se definen como polinomios $P_n(\lambda)$ de los valores propios correspondientes del operador A, mientras que las funciones propias son las mismas que tiene el operador A. Siendo el operador $P_n(A)$ autoconjugado, su norma es igual a un valor propio cuyo módulo es máximo

$$\|P_n(A)\| = \max_{1 \leqslant s \leqslant N} P_n(\lambda_s).$$

Los valores propios λ_s del operador A se disponen en el segmento $[\gamma_1, \gamma_2]$: $\gamma_1 \leqslant \lambda_s \leqslant \gamma_2$. Es evidente que

$$\max_{1 \leqslant s \leqslant N} |P_n(\lambda_s)| \leqslant \max_{\gamma_1 \leqslant x \leqslant \gamma_2} |P_n(x)|,$$

donde el argumento continuo x toma todos los valores en el segmento $[\gamma_1, \gamma_2]$, y, por consiguiente, el problema del mínimo de $\| P_n(A) \|$ se reduce al de míni-máx del polinomio $P_n(x)$, es decir, al problema de determinar $\min_{(\tau_k)} \max_{\gamma \leqslant x_1 \leqslant \gamma_2} | P_n(x) |$.

Apliquemos el segmento $[\gamma_1, \gamma_2]$ sobre el segmento $[-1, 1]$, suponiendo

$$x = \frac{1}{2}[(\gamma_2 - \gamma_1) t + \gamma_2 + \gamma_1], \quad -1 \leqslant t \leqslant 1, \quad \text{para } \gamma_1 \leqslant x \leqslant \gamma_2. \tag{14}$$

Entonces, $P_n(t) = \widetilde{P}_n(t)$. La condición de normalización $P_n(0) = 1$ toma la forma

$$\widetilde{P}_n(t_0) = 1, \qquad t_0 = -1/\rho_0. \tag{15}$$

Así pues, se requiere hallar un polinomio cuya desviación de cero en el segmento $-1 \leqslant t \leqslant 1$ sea mínima, de modo que el $\max | \widetilde{P}_n(t) |$ sea mínimo para la condición complementaria de la normalización (15). El polinomio buscado es

$$\widetilde{P}_n(t) = \frac{T_n(t)}{T_n(t_0)}, \tag{16}$$

donde $T_n(t)$ es el *polinomio de Chébishev*,

$$T_n(t) = \cos(n \arccos t) \quad \text{para} \quad |t| \leqslant 1, \tag{17}$$

$$T_n(t) = \frac{1}{2}[(t + \sqrt{t^2 - 1})^n + ((t - \sqrt{t^2 - 1})^n] \quad \text{para } |t| > 1. \tag{18}$$

El polinomio de Chébishev tiene ceros

$$t_i = \cos\frac{2i-1}{2n}\pi, \qquad i = 1, 2, \ldots, n. \tag{19}$$

El polinomio $P_n(x) = (1 - \tau_1 x)(1 - \tau_2 x) \ldots (1 - \tau_n x)$ tiene ceros $x_i = 1/\tau_i$.

Exigiendo que las raíces de estos polinomios coincidan y teniendo presente la relación (14) entre x y t, obtenemos $2 = [(\gamma_1 + \gamma_2) + (\gamma_2 - \gamma_1) t_i] \tau_i$, de donde se infiere

$$\tau_i = 2/[\gamma_2 + \gamma_1 + (\gamma_2 - \gamma_1) t_i], \quad i = 1, 2, \ldots, n. \quad (20)$$

Esta fórmula queda en vigor, cualquiera que sea el método de poner en orden los ceros del polinomio de Chébishev; por ejemplo, en lugar de (19) podemos hacer $t_i = -\cos \frac{2i-1}{2n} \pi$. Al tener esto en cuenta llegamos a la fórmula (9). Se ha de notar que si $n = 1$, obtenemos $\tau_i = \tau_0$ que es un parámetro optimal del método de iteración simple.

Así pues, los parámetros $\tau_1, \tau_2, \ldots, \tau_n$ se han determinado según (9). Hallemos ahora

$$q_n = \max_{\gamma_1 \leqslant x \leqslant \gamma_2} |P_n(x)| = \max_{-1 \leqslant t \leqslant 1} |\widetilde{P}_n(t)| = \max_{-1 \leqslant t \leqslant 1} \left| \frac{T_n(t)}{T_n(t_0)} \right| = \frac{1}{|T_n(t_0)|},$$

puesto que $\max_{-1 \leqslant t \leqslant 1} |T_n(t)| = 1$. Tenemos $|t_0| > 1$; por *ello, para* $|T_n(t_0)|$ *se usará la fórmula* (18) *con* $t = t_0$. Transformemos las expresiones que figuran en esta fórmula:

$$|t_0| \pm \sqrt{t_0^2 - 1} = \frac{1}{\rho_0} \pm \sqrt{\frac{1}{\rho_0^2} - 1} = \frac{1}{\rho_0} [1 \pm \sqrt{1 - \rho_0^2}] =$$
$$= \frac{1}{\rho_0} \left(1 \pm \frac{2\sqrt{\xi}}{1 - \xi} \right) = \frac{1}{\rho_0} (1 \pm \sqrt{\xi})^2/(1 - \xi) =$$
$$= (1 \pm \sqrt{\xi})^2/(1 - \xi) = (1 \pm \sqrt{\xi})/(1 \mp \sqrt{\xi}),$$

de modo que $|t_0| + \sqrt{t_0^2 - 1} = \frac{1}{\rho_1}$, $|t_0| - \sqrt{t_0^2 - 1} = \rho_1$, y

$$|T_n(t_0)| = \frac{1}{2} \left(\frac{1}{\rho_1^n} + \rho_1^n \right) = \frac{1 + \rho_1^{2n}}{2\rho_1^n} = \frac{1}{q_n}.$$

La estimación (10) queda demostrada.

3. Estabilidad computacional y ordenación de los parámetros. El método iterativo (8) con parámetros de Chébishev $\{\tau_k\}$ se denomina a veces *método de Richardson*. Se conoce desde hace tiempo, no obstante casi no se empleaba en la práctica hasta el último tiempo por su *inestabilidad*

computacional. Expliquemos esta noción con un ejemplo. Tomemos un *sistema de ecuaciones*

$$u(i-1)-2u(i)+u(i+1)=0,\quad i=1,2,\ldots,N-1,$$
$$u(0)=1,\qquad u(N)=0. \tag{21}$$

Su solución es $u(i)=1-x_i$, $x_i=ih$, $h=1/N$. Buscaremos la solución de este problema usando el método iterativo de Chébishev para $N=20$. El valor de $n_0(\varepsilon)$ podemos calcular. Puede resultar no entero. Eligimos el número entero próximo $n \geqslant n_0$. Para N y ε dados tenemos $n(\varepsilon)=64$. Al conocer

$$\gamma_1=\frac{4}{h^2}\operatorname{sen}^2\frac{\pi h}{2},\qquad \gamma_2=\frac{4}{h^2}\cos^2\frac{\pi h}{2},$$
$$h=\frac{1}{N},\qquad \xi=\operatorname{tg}^2\frac{\pi h}{2}\approx\frac{\pi^2h^2}{4}\approx 0{,}006,$$

se puede calcular τ_k según la fórmula (20). A título de la aproximación inicial se toma una función

$$y_0^{(i)}=\begin{cases}1, & i=0.\\ 0, & i>0,\end{cases}$$

Resulta que para el método (8), (9) no es igual en que orden se toman los ceros μ_k del polinomio de Chébishev. He aquí *dos modos de numerar los ceros:*

$$\alpha_1)\ \mu_k=\cos\frac{2k-1}{2n}\pi,\qquad k=1,\ 2,\ \ldots,\ n,$$
$$t_1=\cos\frac{\pi}{2n},\qquad t_n=-\cos\frac{\pi}{2n},$$
$$\alpha_2)\ \mu_k=-\cos\frac{2k-1}{2n}\pi.$$

Los resultados de los cálculos se aducen en la tabla 1.

Para los menores valores de N y n puede resultar que el aumento de los valores intermedios de y_k no lleva al parem, sin embargo tiene lugar la acumulación de los errores de redondeo, y tras n iteraciones no se cumplen las condiciones en que terminan las iteraciones ($\|Ay_k-f\|\leqslant\varepsilon\|Ay_0-f\|$).

Estas dos peculiaridades del proceso computacional, a saber, el aumento de los valores intermedios que conduce al parem y la acumulación de los errores de redondeo, se carac-

TABLA 1

	Surtido α_1		Surtido α_2
k	$\Delta_k = \max\limits_{x_l} \mid y_k(x_l) - y_{k-1}(x_l) \mid$	k	$\Delta_k = \max\limits_{x_l} y_k(x_l) - y_{k-1}(x_l)$
53	0,12	1	39,6
55	27	2	$2{,}6\cdot 10^3$
57	$1{,}9\cdot 10^4$	4	$8{,}2\cdot 10^8$
59	$3{,}7\cdot 10^7$	7	$3{,}3\cdot 10^{11}$
60	$2{,}6\cdot 10^9$	9	$1{,}2\cdot 10^{14}$
61	$2{,}5\cdot 10^{11}$	11	$1{,}9\cdot 10^{16}$
62	$3{,}3\cdot 10^{13}$	12	Parem
63	$5\cdot 10^{15}$		
64	Parem		

terizan por un término: ***inestabilidad computacional***. La causa de inestabilidad computacional del método de Chébishev radica en que las normas $\| S_{k+1} \|$ del operador de paso $S_{k+1} = E - \tau_{k+1}A$ son para ciertas iteraciones superiores a uno, mientras que el proceso computacional es real, es decir, se tienen acotaciones por debajo y por arriba para los números admisibles (se tienen cero de máquina e infinidad de máquina) y en cada etapa de los cálculos surgen los errores de redondeo.

Calculemos la norma para $S_k = E - \tau_k A$. Por cuanto $S_k^* = S_k$, entonces $\| S_{k+1} \| = \sup\limits_{\|x\|=1} |(S_{k+1}x, x)|$. De la condición $\gamma_1 E \leqslant A \leqslant \gamma_2 E$ proviene $(\tau_{k+1}\gamma_1 - 1) E \leqslant \leqslant \tau_{k+1}A - E \leqslant (\tau_{k+1}\gamma_2 - 1) E$. Sustituyendo aquí la expresión para τ_{k+1} y teniendo presente que $1 - \tau_0\gamma_1 = = \tau_0\gamma_2 - 1 = \rho_0$, obtenemos

$$-\frac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k} E \leqslant \tau_{k+1}A - E \leqslant \frac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k} E.$$

De aquí encontramos

$$\|S_{k+1}\| = \|\tau_{k+1}A - E\| = \begin{cases} \dfrac{\rho_0(1+\mu_k)}{1+\rho_0\mu_k} & \text{para } \mu_k > 0, \\[2mm] \dfrac{\rho_0(1-\mu_k)}{1+\rho_0\mu_k} & \text{para } \mu_k < 0 \end{cases}$$

de suerte que $\|S_{k+1}\| < 1$ para todos los $\mu_k > 0$, y $\|S_{k+1}\| > 1$ para $\mu_k < -(1-\rho_0)/(2\rho_0)$. Por cuanto

$$-\cos\frac{\pi}{2n} \leqslant \mu_k \leqslant -\cos\frac{(2n-1)}{2n}\pi = \cos\frac{\pi}{2n},$$
$$k = 1, 2, \ldots, n,$$

entonces, para la mayor cantidad de números k tenemos $\| S_k \| > 1$, y si se emplean muchos parámetros τ_k seguidos, para los cuales $\| S_k \| > 1$, tienen lugar la acumulación del error de redondeo y el crecimiento de las aproximaciones iterativas, lo que conduce a la inestabilidad computacional.

Con el fin de debilitar el efecto mencionado, es natural tratar de alternar los parametros τ_k, para los cuales $\| S_k \| > 1$ con aquellos, para los cuales $\| S_k \| < 1$. En este procedimiento se realiza precisamente la construcción de una sucesión de parámetros $\{\tau_k\}$, para la cual la convergencia de las iteraciones tiene un carácter monótono y la inestabilidad computacional está ausente. Existe una regla para tal ordenación de los ceros $t_i = -\cos\frac{2i-1}{2n}\cos\pi$ del polinomio de Chébishev y, consecuentemente, también de los parámetros $\{\tau_k\}$ (para n cualquiera), para la que tiene lugar la estabilidad computacional.

Demos a conocer dicha regla para el caso en que n es una potencia del número 2, $n = 2^p$, $p > 0$ es un número entero [1]). Designemos el conjunto de ceros t_i, ordenado según esta regla, mediante

$$\mathfrak{M}_n^* = \left\{ -\cos\beta_i,\ \beta_i = \frac{\pi}{2n}\theta_i^{(n)},\ i = 1, 2, \ldots, n \right\},\quad n = 2^p,$$

donde $\theta_i^{(n)}$ es uno de los números impares $1, 3, 5, \ldots, 2n-1$. El problema se reduce, pues, a la ordenación del conjunto de n números impares: $\theta_n = \{\theta_1^{(n)}, \theta_2^{(n)}, \ldots, \theta_n^{(n)}\}$. Partiendo del conjunto $\theta_1 = \{1\}$, construyamos un conjunto $\theta_{(n)}^* = \theta_{2^p}^*$ según las fórmulas

$$\theta_{2i-1}^{(2m)} = \theta_i^{(m)},$$
$$\theta_{2i}^{(2m)} = 4m - \theta_{2i-1}^{(2m)},\quad i = 1, 2, \ldots, m;\ m = 1, 2, \ldots, 2^{p-1},$$

[1]) La regla de ordenación de $\{\tau_k\}$ para n cualquiera se da, por ejemplo, en [6, 9].

si se conocen $\theta_i^{(m)}$. La sucesión correspondiente de parámetros $\{\tau_k^*\}$ se denominará *surtido estable*. Sea, por ejemplo, $n = 16 = 2^4$. Encontramos sucesivamente $\theta_1 = \{1\}$, $\theta_2 = \{1, 3\}$, $\theta_4 = \{1, 7, 3, 5\}$, $\theta_8 = \{1, 15, 7, 9, 3, 13, 5, 11\}$, $\theta_{16} = \{1, 31, 15, 17, 7, 25, 9, 23, 3, 29, 13, 19, 5, 27, 11, 21\}$. Al pasar de θ_m a θ_{2m} es suficiente poner tras cada $\theta_{2i-1}^{(m)}$ un número igual a $4m - \theta_{2i-1}^{(m)}$ (la numeración corresponde a θ_m). La sucesión «estable» θ_n^* no depende del problema. La convergencia de las iteraciones para este surtido de parámetros $\{\tau_k^*\}$ lleva un carácter no monótono, pero las oscilaciones aquí no son de gran amplitud y se amortiguan al fin y al cabo.

He aquí los resultados de cálculos para el problema (21) según el esquema (8), (9) con el surtido estable de parámetros $\{\tau_k^*\}$:

k	1	4	8	16	24	32	48	50	62
Δk	39,6	4,7	1,1	0,2	0,1	0,04	$1{,}5\cdot 10^{-3}$	$6{,}7\cdot 10^{-4}$	$8{,}7\cdot 10^{-5}$

4. **Esquemas implícitos.** El método de Seidel y el de relajación superior convergen más rápidamente que el método explícito de iteración simple, razón por la cual se justifica *el paso a los esquemas implícitos. ¿Cómo se debe elegir el* operador B? Es fundamental el requisito general del mínimo de operaciones $Q(\varepsilon)$ necesarias para hallar solución con una exactitud $\varepsilon > 0$; dicho requisito se reduce a dos exigencias: 1) del número mínimo de iteraciones, el cual depende tanto de B como de la elección de $\{\tau_k\}$; 2) del número mínimo de operaciones para la resolución de la ecuación

$$By_{k+1} = F_k$$

(*carácter económico* del operador B). De ejemplo puede servir un operador triangular correspondiente a la matriz triangular.

Mostremos ahora que los resultados obtenidos más arriba para un esquema explícito pueden extenderse a un esquema implícito. Veamos un esquema implícito

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, \text{ para todo } y_0 \in H, \tag{22}$$

donde $A = A^* > 0$, $B = B^* > 0$, y

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \qquad \gamma_1 > 0. \tag{23}$$

Eligiendo los parámetros de iteración $\{\tau_k^*\}$ según las fórmulas (9) y ordenándolos en concordancia con el punto precedente, obtendremos para la solución del problema (22) una estimación

$$\|Ay_n - f\|_{B^{-1}} \leqslant q_n \|Ay_0 - f\|_{B^{-1}}, \qquad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}},$$

$$\rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \qquad \xi = \frac{\gamma_1}{\gamma_2}, \qquad (24)$$

donde γ_1 y γ_2 son los números que figuran en (23). Para el número de iteraciones $n = n(\varepsilon)$ son válidas las estimaciones (11) y (12). Para convencernos de esto, basta reducir el problema (22) a un problema equivalente para el esquema explícito

$$\frac{x_{k+1} - x_k}{\tau_{k+1}} + Cx_k = 0, \qquad k = 0, 1, \ldots, x_0 = B^{1/2}w_0, \qquad (25)$$

donde $x_k = B^{1/2}w_k$, $C = B^{-1/2}AB^{-1/2}$ es un operador positivo autoconjugado con las fronteras del espectro γ_1 y γ_2:

$$\gamma_1 E \leqslant C \leqslant \gamma_2 E. \qquad (26)$$

En efecto, por cuanto $B = B^* > 0$, existe, pues, $B^{1/2} = (B^{1/2})^* > 0$. Aplicando el operador $B^{-1/2}$ a la ecuación (22), obtenemos (25) para $x_k = B^{1/2}w_k$. El modo inverso de razonamientos es evidente. Queda por demostrar la equivalencia de las desigualdades (23) y (26). Estudiemos una funcional

$$J = ((A - \gamma B)\, y, y) = (Ay, y) - \gamma\,(By, y) =$$
$$= (AB^{-1/2}(B^{1/2}y), B^{-1/2}(B^{1/2}y)) - \gamma\,(B^{1/2}y, B^{1/2}y) =$$
$$= (Cx, x) - \gamma\,(x, x) = ((C - \gamma E)\, x, x),$$

donde $x = B^{1/2}y$. Como y (y, por tanto, también x) es un vector arbitrario de H, entonces de la igualdad

$$J = ((A - \gamma B)\, y, y) = ((C - \gamma E)\, x, x) \qquad (27)$$

se deduce que los operadores $A - \gamma B$ y $C - \gamma E$ son de signos iguales. Si, por ejemplo, $A - \gamma_1 B \geqslant 0$, entonces para $\gamma = \gamma_1$ la igualdad (27) nos da $C - \gamma_1 E \geqslant 0$, etc.

Para el esquema explícito tenemos una estimación $\| x_n \| \leqslant q_n \| x_0 \|$. Al sustituir aquí $x_k = B^{1/2}w_k = B^{-1/2}r_k$, $r_k = Ay_k - f$, obtenemos la estimación (24).

Para los metodos de Seidel y de relajación superior $B \neq B^*$, *por lo cual la totalidad de parámetros de Chébishev no puede ser empleado.*

§ 5. Método alternado triangular

1. Método alternado triangular. Analizaremos un esquema iterativo implícito

$$B\frac{y_{k+1}-y_k}{\tau_{k+1}} + Ay_k = f, \qquad k=0,\ 1,\ \ldots \tag{1}$$

Si el operador B representa un producto del número finito de operadores económicos, será también económico. Así, es económico el operador $B = B_1 B_2$ que es igual al producto de los operadores triangulares B_1 y B_2.

Veamos el así llamado método *alternado triangular* (1), *para el cual el operador B tiene por expresión*

$$B = (D + \omega R_1)\, D^{-1}\, (D + \omega R_2), \tag{2}$$

donde $D = D^* > 0$, $R_1^* = R_2$, $R_1 + R_2 = R$, $R = R^* > > 0$, $\omega > 0$ es el parámetro.

Probemos que el operador B es positivo y autoconjugado, es decir, que el esquema (1) con el operador (2) pertenece *a la familia inicial de esquemas* (2) del § 3, razón por la cual pueden aprovecharse todos los resultados de la teoría general obtenidos anteriormente. En efecto,

$$\begin{aligned}(By,\ v) &= ((D + \omega R_1)\, D^{-1}\, (D + \omega R_2)\, y,\ v) = \\ &= ((D + \omega R_2)\, y,\ D^{-1}\, (D + \omega R_2)\, v) = \\ &= (y,\ (D + \omega R_1)\, D^{-1}\, (D + \omega R_2)\, v),\end{aligned}$$

y, por consiguiente, $(By,\ v) = (y,\ Bv)$, es decir, $B = B^*$. Luego, encontramos $(By,\ y) = ((D + \omega R_2)\, y,\ D^{-1} \times \times D^{-1}\, (D + \omega R_2)\, y) = \| (D + \omega R_2)\, y \|^2_{D^{-1}} > 0$, es decir, $B = B^* > 0$.

Al operador R le corresponde una matriz $R = (r_{ij})$. A título de las matrices R_1 y R_2 pueden intervenir las matri-

ces triangulares inferior y superior, es decir,

$$R_1=(r_{ij}^-), \qquad r_{ij}^-=\begin{cases} r_{ii}/2, & j=i, \\ r_{ij}, & j<i, \\ 0, & j>i; \end{cases}$$

$$R_2=(r_{ij}^+), \qquad r_{ij}^+=\begin{cases} r_{ii}/2, & j=i, \\ r_{ij}, & j>i, \\ 0, & j<i. \end{cases}$$

Si R es una matriz simétrica, $r_{ji} = r_{ij}$, entonces R_1 y R_2 son recíprocamente conjugados, $R_2 = R_1^*$.

A título de $D = (d_{ij})$ tomemos una matriz diagonal. Entonces, $D + \omega R_1$ es una matriz triangular inferior y $D + \omega R_2$, una matriz triangular superior. De este modo, el proceso de la iteración se reduce a la inversión alternada de las matrices triangulares inferior y superior (de aquí proviene la denominación del método). Efectivamente, para cada iteración se debe resolver una ecuación

$$By_{k+1} = (D + \omega R_1)\, D^{-1}\, (D + \omega R_2)\, y_{k+1} = F_k. \tag{3}$$

Al denotar $D^{-1}(D + \omega R_2)\, y_{k+1} = \bar{y}_{k+1}$, obtenemos

$$(D + \omega R_1)\, \bar{y}_{k+1} = F_k, \qquad (D + \omega R_2)\, y_{k+1} = D\bar{y}_{k+1}, \qquad k = 0, 1, \ldots \tag{4}$$

Observando que $(R_1y, y) = (R_2y, y) = (Ry, y)/2$, encontramos

$$((D+\omega R_1)\, y, y) = (Dy, y) + \omega\,(R_1y, y) = \left(\left(D+\frac{\omega}{2}R\right) y, y\right) = ((D+\omega R_2)\, y, y) > 0,$$

puesto que $D > 0$, $\omega > 0$, y $R > 0$.

De aquí proviene la existencia de los operadores inversos $(D + \omega R_1)^{-1}$, $(D + \omega R_2)^{-1}$, es decir, la resolubilidad de los problemas (4).

2. Elección del parámetro ω. Para poder emplear la teoría general, se deben hallar primeramente los parámetros γ_1 y γ_2 que figuran en las desigualdades

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \tag{5}$$

las cuales siempre se verifican gracias a que los operadores A y B son acotados y positivos. Empecemos con la determinación del parámetro $\omega > 0$.

LEMA. *Supongamos que el operador B se determina según la fórmula* (2), *donde*

$$R_2^* = R_1, \qquad R_1 + R_2 = R, \qquad R = R^* > 0$$

y que R satisface las condiciones

$$R \geqslant \delta D, \qquad \delta > 0, \qquad R_1 D^{-1} R_2 \leqslant \frac{\Delta}{4} R, \qquad \Delta > 0, \qquad (6)$$

En este caso será válida la estimación

$$\overset{\circ}{\gamma}_1 B \leqslant R \leqslant \overset{\circ}{\gamma}_2 B, \quad \overset{\circ}{\gamma}_1 = \frac{\delta}{1 + \omega\delta + 0{,}25\omega^2\delta\Delta}, \quad \overset{\circ}{\gamma}_2 = \frac{1}{2\omega}. \qquad (7)$$

La razón $\xi = \overset{\circ}{\gamma}_1(\omega)/\overset{\circ}{\gamma}_2(\omega)$ tiene el valor máximo cuando

$$\omega = \overset{\circ}{\omega} = 2/\sqrt{\delta\Delta};$$

con la particularidad de que

$$\xi = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \quad \eta = \frac{\delta}{\Delta}, \quad \overset{\circ}{\gamma}_1 = \frac{\delta}{2(1+\sqrt{\eta})}, \quad \overset{\circ}{\gamma}_2 = \frac{\delta}{4\sqrt{\eta}}. \quad (9)$$

DEMOSTRACION. Las desigualdades (6) significan que

$$(Ry, y) \geqslant \delta (Dy, y), \quad (D^{-1}R_2 y, R_2 y) \leqslant \frac{\Delta}{4} (Ry, y)$$

para cualesquiera $y \in H$.

Realizadas las transformaciones

$$B = (D + \omega R_1) D^{-1} (D + \omega R_2) =$$
$$= D - \omega (R_1 + R_2) + \omega^2 R_1 D^{-1} R_2 + 2\omega (R_1 + R_2) =$$
$$= (D - \omega R_1) D^{-1} (D - \omega R_2) + 2\omega R,$$

obtenemos

$$(By, y) = (D^{-1} (D - \omega R_2) y, (D - \omega R_2) y) + 2\omega (Ry, y)$$
$$= \| (D - \omega R_2) y \|^2_{D^{-1}} + 2\omega (Ry, y) \geqslant 2\omega (Ry, y),$$

de suerte que

$$B \geqslant 2\omega R, \text{ o bien } R \leqslant \frac{1}{2\omega} B, \qquad \overset{\circ}{\gamma}_2 = \frac{1}{2\omega}.$$

Obtendremos ahora para B la estimación por arriba. Tomando en consideración (6), hallemos

$$B = D + \omega R + \omega^2 R_1 D^{-1} R_2 \frac{1}{\delta} R + \omega R + \frac{\omega^2 \Delta}{4} R \leqslant$$
$$\leqslant \frac{1}{\delta}\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right) R,$$
$$R_1 \geqslant \mathring{\gamma}_1 B, \qquad \mathring{\gamma}_1 = \delta\left(1 + \omega\delta + \frac{\omega^2\delta\Delta}{4}\right)^{-1}.$$

El número de iteraciones necesarias para la resolución de la ecuación $Ry = f$ depende de la razón

$$\mathring{\xi}(\omega) = \mathring{\gamma}_1/\mathring{\gamma}_2 = 2\omega\delta\,(1 + \omega\delta + \omega^2\delta\Delta/4)^{-1}.$$

Elijamos ω de la condición de que $\mathring{\xi}(\omega)$ sea máxima. Igualando a cero la derivada $\mathring{\xi}'(\omega) = 2\delta\,(1 - \omega^2\delta\Delta/4)\,(1 + \omega\delta + \omega^2\delta\Delta/4)^{-2}$, encontramos $\omega = \mathring{\omega} = 2/\sqrt{\delta\Delta}$; en este caso $\mathring{\xi}''(\mathring{\omega}) < 0$. Al sustituir este valor de ω en las fórmulas para $\mathring{\gamma}_1$, $\mathring{\gamma}_2$, $\mathring{\xi}(\omega)$, obtenemos la fórmula (9). El lema queda demostrado.

3. Velocidad de convergencia.

TEOREMA. *Supongamos que el operador* $A = A^* > 0$ *se representa como una suma* $A = A_1 + A_2$, $A_2 = A_1^*$, *y que se cumplen las condiciones*

$$A \geqslant \delta D, \quad A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta > 0. \qquad (10)$$

Entonces para el método alternado triangular (1) *con*

$$B = (D + \omega A_1)\, D^{-1}\, (D + \omega A_2), \quad D = D^* > 0, \qquad (11)$$

con el parámetro $\omega = 2/\sqrt{\delta\Delta}$ *y una totalidad de parámetros de Chébishev*

$$\tau_k^* = \frac{\tau_0}{1 + \rho_0 \mu_k^*}, \qquad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}, \qquad \rho_0 = \frac{1-\xi}{1+\xi},$$
$$\xi = \frac{\gamma_1}{\gamma_2} = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}} \qquad (12)$$

donde

$$\gamma_1=\frac{\delta}{2(1+\sqrt{\eta})},\quad \gamma_2=\frac{\delta}{4\sqrt{\eta}},\quad \eta=\frac{\delta}{\Delta},\quad \mu_k^*\in\overline{\mathfrak{M}}_n^*, \tag{13}$$

son suficientes $n(\varepsilon)$ iteraciones:

$$n_0(\varepsilon)\leqslant n(\varepsilon)<n_0(\varepsilon)+1,\quad n_0(\varepsilon)<\ln\frac{2}{\varepsilon}\Big/(2\sqrt{2}\sqrt[4]{\eta}), \tag{14}$$

y en este caso se cumple la siguiente estimación

$$\| Ay_n-f\|_{B^{-1}}\leqslant\varepsilon\| Ay_0-f\|_{B^{-1}}. \tag{15}$$

DEMOSTRACIÓN. Hagamos uso del lema precedente suponiendo $R=A$, $R_1=A_1$, $R_2=A_2$, y también de la estimación (24) del § 4:

$$\| Ay_n-f\|_{B^{-1}}\leqslant q_n\| Ay_0-f\|_{B^{-1}}$$

con

$$q_n=\frac{2\rho_1^n}{1+\rho_1^{2n}},\qquad \rho_1=\frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

En el § 3 se ha obtenido la estimación para el número de iteraciones $n=n(\varepsilon)$:

$$n_0(\varepsilon)\leqslant n(\varepsilon)<n_0(\varepsilon)+1,\ \text{donde}\ n_0(\varepsilon)<\ln\frac{2}{\varepsilon}\Big/(2\sqrt{\xi}).$$

Al sustituir aquí $\xi=2\sqrt{\eta}/(1+\sqrt{\eta})$, obtendremos (15).

4. Ejemplo de aplicación del método alternado triangular. Analicemos un problema modelo

$$u_{\bar{x}x,i}=\frac{u_{i-1}-2u_i+u_{i+1}}{h^2}=-f_i,\quad i=1,2,\ldots,N-1,$$
$$u_0=0,\qquad u_N=0. \tag{16}$$

Sea $H=\Omega$ un espacio de funciones reticulares definidas en los nodos interiores $i=1,2,\ldots,N-1$ de la red ω_h; introduzcamos un producto escalar

$$(y,v)=\sum_{i=1}^{N-1}y_iv_ih.$$

El operador $Ay = -\bar{y}_{xx}$ es autoconjugado y definido positivo:

$$A \geqslant \delta E, \qquad \delta = \frac{4}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}.$$

Introduzcamos los operadores $Dy = y$ $(D = E)$ y

$$A_1 y = R_1 y = \frac{y_{\bar{x}, i}}{h} = \frac{y_i - y_{i-1}}{h^2},$$

$$A_2 y = R_2 y = -\frac{y_{x, i}}{h} = -\frac{y_{i+1} - y_i}{h^2}, \quad A_1 + A_2 = A.$$

Las iteraciones $(y_i)_k = y_k(i)$ se hallan según las fórmulas

$$(E + \omega A_1)(\bar{y}_i)_{k+1} = \left(\bar{y}_i + \omega \frac{\bar{y}_i - \bar{y}_{i-1}}{h^2}\right)_{k+1} = F_k(i),$$

$$\bar{y}_{k+1}(i) = \frac{\omega \bar{y}_{k+1}(i-1) + h^2 F_k(i)}{h^2 + \omega},$$

$$(E + \omega A_2) y_{k+1}(i) =$$
$$= y_{k+1}(i) - \frac{\omega}{h^2}(y_{k+1}(i+1) - y_{k+1}(i)) = \bar{y}_{k+1}(i),$$

en definitiva tenemos

$$y_{k+1}(i) = \frac{\omega y_{k+1}(i+1) + h^2 \bar{y}_{k+1}(i)}{\omega + h^2},$$
$$i = N-1, \ N-2, \ \ldots, \ 2, \ 1.$$

Los valores de $\bar{y}_{k+1}(i)$ se hallan sucesivamente al mover de izquierda a derecha (de $i - 1$ a i) y los de $y_{k+1}(i)$, de derecha a izquierda (de $i + 1$ a i); y en este caso se toman en consideración las condiciones de contorno

$$\bar{y}_{k+1}(0) = 0, \qquad y_{k+1}(N) = 0.$$

Las fórmulas del tipo semejante se denominan *fórmulas de cómputo móvil.*

De las igualdades $y_{\bar{x}, i+1} = y_{x, i}$ se desprende que $A_1^* = A_2$. En efecto, por cuanto $v_1 = v_0 + hv_{\bar{x},1} = hv_{\bar{x},1}$,

entonces

$$(A_2y, v) = -\sum_{i=1}^{N-1} y_{x,i}v_i = -y_1v_1\frac{1}{h} - \sum_{i=1}^{N-1} y_{x+1}x_{x,i} =$$

$$= y_1v_{\bar{x},1} + \sum_{i=2}^{N} y_iv_{\bar{x},i} = \sum_{i=1}^{N-1} y_iv_{\bar{x},i} = h\sum_{i=1}^{N-1} y_i\frac{v_{\bar{x},i}}{h} = (y, A_1v),$$

es decir, $A_1 = A_2^*$.

Calculemos la constante Δ:

$$(A_1A_2y, y) = (A_2y, A_2y) =$$

$$= \frac{1}{h^2}\sum_{i=1}^{N-1}(y_{x,i})^2h = \frac{1}{h^2}\sum_{i=2}^{N-1}(y_{\bar{x},i})^2h \leqslant$$

$$\leqslant \frac{1}{h^2}\sum_{i=1}^{N} h(y_{\bar{x},i})^2 = \frac{1}{h^2}\sum_{i=1}^{N-1} h(Ay)_iy_i = \frac{1}{h^2}(Ay, y),$$

de donde se infiere $\Delta = 4/h^2$. Así que,

$$\eta = \frac{\delta}{\Delta} = \operatorname{sen}^2\frac{\pi h}{2} \approx \frac{\pi^2h^2}{4}, \quad \sqrt{\eta} \approx \frac{\pi h}{2},$$

$$\xi = 2\sqrt{\eta}/(1+\sqrt{\eta}) \approx 2\sqrt{\eta} \approx \pi h, \quad \sqrt{\xi} \approx \sqrt{\pi h},$$

de suerte que

$$n_0(\varepsilon)\frac{1}{2\sqrt{\pi h}}\ln\frac{2}{\varepsilon}.$$

Si $\varepsilon = 10^{-4}$, entonces $n_0(\varepsilon) \approx 3/\sqrt{h}$.

El resultado es:

$$n_0(\varepsilon) \approx 10 \text{ para } h = 1/10 \ (N = 10),$$

$$n_0(\varepsilon) \approx 30 \text{ para } h = 1/100 \ (N = 100).$$

Recordemos que para $N = 100$ se deben realizar 20 000 iteraciones por el método de iteración simple y 340 iteraciones, por el esquema explícito de Chébishev. De este modo, el método alternado triangular ha resultado mejor entre los métodos estudiados.

§ 6. Métodos iterativos de tipo variacional

1. Método de los residuos mínimos. Al estudiar los métodos iterativos siempre se suponía hasta ahora que las constantes γ_1 y γ_2, es decir, las fronteras del espectro del operador A en H o en H_B están conocidas. Pero, ¿qué se debe hacer, si tal información está ausente? En este caso pueden emplearse los métodos que no utilizan los parámetros γ_1 y γ_2 en la forma explícita. Estos son los métodos de tipo variacional. Aquí se analizarán los métodos de los residuos mínimos, del descenso más rápido y de los gradientes conjugados.

Empecemos con el *método de los residuos mínimos* para un esquema explícito

$$\frac{y_{k+1}-y_k}{\tau_{k+1}}+Ay_k=f,\quad k=0,\ 1,\ \ldots,\quad \text{Para todo } y_0\in H. \quad (1)$$

Para el residuo $r_k = Ay_k - f$ tenemos una ecuación homogénea

$$\frac{r_{k+1}-r_k}{\tau_{k+1}}+Ar_k=0,\quad k=0,\ 1,\ \ldots,\quad r_0=Ay_0-f. \quad (2)$$

El parámetro τ_{k+1} se escogerá partiendo de la condición de que sea mínimo el residuo r_{k+1} según la norma:

$$\|r_{k+1}\|^2=\|r_k-\tau_{k+1}Ar_k\|^2=$$
$$=\|r_k\|^2-2\tau_{k+1}(r_k,\ Ar_k)+\tau_{k+1}^2\|Ar_k\|^2=\varphi(\tau_{k+1})$$

Diferenciemos esta expresión respecto de τ_{k+1}, igualemos a cero la derivada $\varphi'(\tau_{k+1})$:

$$\varphi'(\tau_{k+1}) = -2(r_k,\ Ar_k) + 2\tau_{k+1}\|Ar_k\|^2 = 0$$

y hallemos

$$\tau_{k+1}=\frac{(Ar_k,\ r_k)}{\|Ar_k\|^2},\qquad k=1,\ 2,\ \ldots \quad (3)$$

Con este valor de τ_{k+1} la segunda derivada $\varphi''(\tau_{k+1})$ es positiva y, por consiguiente, se alcanza el $\min\limits_{\tau_{k+1}}\|r_{k+1}\|^2$.

Hasta ahora no se suponía que A es un operador autoconjugado. En cambio, si $A = A^* > 0$, entonces son válidas

las estimaciones

$$\| r_{k+1} \| \leqslant \rho_0 \| r_k \|, \qquad \| Ay_n - f \| \leqslant \rho_0^n \| Ay_0 - f \|,$$
$$\rho_0 = \frac{1-\xi}{1+\xi}, \qquad \xi = \frac{\gamma_1}{\gamma_2}, \tag{4}$$

donde γ_1 y γ_2 son las fronteras exactas del espectro del operador A. En efecto, por cuanto, de acuerdo con (3), la norma $\| r_{k+1} \|$ es mínima para τ_{k+1}, entonces para cada $\tau \neq \tau_{k+1}$ ella debe crecer, por lo cual

$$\| r_{k+1} \|^2 = \| r_k - \tau_{k+1} A r_k \|^2 \leqslant \| r_k - \tau_0 A r_k \|^2 \leqslant \\ \leqslant \| E - \tau_0 A \|^2 \| r_k \|^2.$$

Por otra parte se conoce que

$$\| E - \tau_0 A \| = \rho_0 \text{ para } \tau_0 = 2/(\gamma_1 + \gamma_2).$$

De aquí precisamente se deduce que $\| r_{k+1} \| \leqslant \rho_0 \| r_k \|$.

De este modo, el método de los residuos mínimos converge con la misma velocidad que el método de iteración simple (siempre que en este último se empleen los valores exactos de γ_1 y γ_2).

En el caso del método de residuos implícito o del método de correcciones en vez de (1) obtenemos una ecuación para la corrección:

$$B \frac{w_{k+1} - w_k}{\tau_{k+1}} + A w_k = 0, \qquad k = 0, 1, \ldots,$$
$$w_k = B^{-1} r_k, \qquad w_0 = B^{-1} (Ay_0 - f), \tag{5}$$

donde τ_{k+1} se determina por la fórmula

$$\tau_{k+1} = \frac{(Aw_k, w_k)}{(B^{-1} A w_k, A w_k)}, \qquad k = 0, 1, \ldots \tag{6}$$

En lugar de (4) obtenemos la estimación

$$\| Ay_n - f \|_{B^{-1}} \leqslant \rho_0^n \| Ay_0 - f \|_{B^{-1}}.$$

2. Método del descenso más rápido. El método explícito ***del descenso más rápido*** se diferencia del método de los residuos mínimos sólo en la fórmula para τ_{k+1}:

$$\tau_{k+1} = \frac{(r_k, r_k)}{(Ar_k, r_k)}, \qquad k = 0, 1, \ldots \tag{7}$$

Esta fórmula se obtiene o bien de la condición del mínimo de la norma $\| z_{k+1} \|_A$ del error $z_{k+1} = y_{k+1} - u$, o bíen de la condición de ortogonalidad de los residuos r_k y $_{k+1}$. Al multiplicar escalarmente la ecuación $r_{k+1} = r_k - \tau_{k+1} A r_k$ por r_k, obtenemos $0 = (r_k, r_k) - \tau_{k+1} (A r_k, r_k)$ de donde se infiere la fórmula (7). Por cuanto $A z_k = A y_k - A u = r_k$, entonces se tiene

$$(A z_{k+1}, z_{k+1}) = (A z_k - \tau_{k+1} A^2 z_k, z_k - \tau_{k+1} A z_k) = \\ = (r_k - \tau_{k+1} A r_k, z_k - \tau_{k+1} r_k) = \\ = (r_k, z_k) - 2\tau_{k+1} (r_k, r_k) + \tau_{k+1}^2 (A r_k, r_k).$$

Diferenciando $\| z_{k+1} \|_A^2$ respecto de τ_{k+1} e igualando a cero la derivada, obtendremos (7).

Luego, tenemos

$$\| z_{k+1} \|_A^2 = \| (E - \tau_{k+1} A) z_k \|_A^2 \leqslant \| (E - \tau_0 A) z_k \|_A^2 \leqslant \\ \leqslant \| E - \tau_0 A \|^2 \| z_k \|_A^2 \leqslant \rho_0^2 \| z_k \|_A,$$

es decir,

$$\| z_{k+1} \|_A = \| y_{k+1} - u \|_A \leqslant \rho_0^n \| y_0 - u \|_A.$$

El método del descenso más rápido converge en H_A con la misma velocidad que el método de iteración simple.

3. Método de los gradientes conjugados. Los métodos de tipo variacional con mayor velocidad de convergencia pueden encontrarse en la clase de esquemas iterativos implícitos de tres capas:

$$B y_{k+1} = \alpha_{k+1} (B - \tau_{k+1} A) y_k + (1 - \alpha_{k+1}) B y_{k-1} + \\ + \alpha_{k+1} \tau_{k+1} f, \quad k = 1, 2, \ldots, \tag{8}$$
$$B y_k = (B - \tau_1 A) y_0 + \tau_1 f.$$

Veamos el *método de gradientes conjugados* que es de amplio uso en la práctica. En este método los parámetros iterativos α_{k+1} y τ_{k+1} se determinan según las fórmulas

$$\tau_{k+1} = \frac{(r_k, w_k)}{(A w_k, w_k)},$$
$$\alpha_{k+1} = \left(1 - \frac{\tau_{k+1}}{\tau_k} \frac{(r_k, w_k)}{(r_{k-1}, w_{k-1})} \frac{1}{\alpha_k} \right)^{-1}, \tag{9}$$

donde $k = 0, 1, 2, \ldots$, bajo el supuesto de que $A = A^* > 0$, $B = B^* > 0$, $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, $\gamma_1 > 0$. Las fórmulas para τ_{k+1}, α_{k+1} se obtienen del requisito del mínimo de la norma del operador de resolución. Con estos valores optimales de los parámetros iterativos queda lícita la estimación

$$\| y_n - u \|_A \leqslant q_n \| y_0 - u \|_A, \qquad q_n = \frac{2\rho_1^n}{1+\rho_1^{2n}},$$
$$\rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}, \qquad \xi = \frac{\gamma_1}{\gamma_2}, \quad (10)$$

es decir, la velocidad de convergencia del método de gradientes conjugados es la misma que la del método iterativo de dos capas con parámetros de Chébishev (en el que se usan γ_1 y γ_2 al calcular los parámetros τ_{k+1}). Por eso, para el número de iteraciones tenemos una estimación

$$n_0(\varepsilon) \leqslant n(\varepsilon) \leqslant n_0(\varepsilon) + 1, \qquad n_0(\varepsilon) = \frac{1}{2\sqrt{\xi}} \ln \frac{2}{\varepsilon}.$$

A título de operador B podemos tomar el operador factorizado del método alternado triangular

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2),$$
$$A_1 + A_2 = A > 0, \quad A_1^* = A_2, \quad D = D^* > 0.$$

Los cálculos muestran que el número de iteraciones al aplicar el método alternado triangular en conjunto con el de los gradientes conjugados es menor que en el caso de emplear el esquema de Chébishev.

§ 7. Resolución de las ecuaciones no lineales

1. Métodos iterativos. Analicemos una ecuación no lineal

$$f(x) = 0, \qquad x \in [a, b],$$

donde $f(x)$ es una función continua. La ecuación puede tener una o varias raíces. Se pide; 1) establecer la existencia de las raíces de la ecuación; 2) hallar los valores aproximados de las raíces. Ambos problemas se resuelven a menudo simultáneamente. *Para hallar las raíces se emplean los métodos iterativos.*

El más elemental es el *método de dicotomía* (división por la mitad). Sea $f(x_0) f(x_1) \leqslant 0$; entonces, en el segmento $[x_0, x_1]$ se ubica por lo menos una raíz. Determinemos $f(x_2)$, donde $x_2 = (x_0 + x_1)/2$, y elijamos x_3: aquél de los valores x_0 o x_1, para el cual se cumple la condición $f(x_2) f(x_3) \leqslant \leqslant 0$. El segmento $[x_2, x_3]$ se dividirá por la mitad de nuevo, etc. La división continúa hasta que la longitud del segmento se haga inferior a 2ε, donde ε es la exactitud con la que se debe determinar la raíz. En este caso el centro de dicho segmento nos presta precisamente el valor de la raíz con la exactitud requerida ε. Es evidente que el proceso converge con la velocidad de una progresión geométrica de razón 1/2. La deficiencia del método consiste en la elección del segmento inicial $[x_0, x_1]$: no está claro de antemano a qué raíz convergerá el proceso (si en $[x_0, x_1]$ hay varias raíces).

El segundo método es el de *iteración simple*. Escribamos la ecuación (1) en la forma

$$x = \varphi(x), \tag{2}$$

donde $\varphi(x)$ puede definirse por uno de los siguientes métodos:

$$\varphi(x) = x - \alpha f(x), \quad \alpha = \text{const},$$

$\varphi(x) = x + \rho(x) f(x)$, $\rho(x)$ es una función arbitraria que no tiene raíces en el segmento $[a, b]$.

El método de iteración simple se determina por la fórmula

$$x_{n+1} = \varphi(x_n), \quad n = 0, 1, 2, \ldots, \tag{3}$$

donde n es el número de la iteración, x_0 es la aproximación inicial prefijada arbitrariamente. Se pide hallar aproximadamente la solución (la raíz) $x = x^*$ de la ecuación $x = = \varphi(x)$ con un error relativo $\varepsilon > 0$ de un modo tal que para cualquier $n \geqslant n_0$ se verifique la desigualdad

$$|x_n - x^*| \leqslant \varepsilon |x_0 - x^*|, \quad n \geqslant n_0(\varepsilon). \tag{4}$$

Esta condición puede cumplirse, siempre que la sucesión de iteraciones $\{x_n\}$ converja, para $n \to \infty$, al límite x^*: : $\lim\limits_{n \to \infty} x_n = x^*$. Si (4) tiene lugar, los cálculos pueden ser terminados con $n = n_0$. De aquí se ve que la cuestión más importante en este caso es la de convergencia de las iteracio-

nes, como también de la velocidad de su convergencia, es decir, la cuestión sobre el número mínimo de iteraciones $n_0(\varepsilon)$, para el cual queda cumplida la desigualdad (4). Supongamos que en cierto δ-entorno

$$\Delta = (x_0 - \delta,\ x_0 + \delta), \qquad \delta > 0, \tag{5}$$

del punto x_0 la función $\varphi(x)$ satisface la condición de Lipschitz:

$$|\varphi(x'') - \varphi(x')| \leqslant q\,|x'' - x'| \text{ para cualesquiera } x',\ x'' \in \Delta \tag{6}$$

con el coeficiente $q < 1$:

$$0 < q < 1 \tag{7}$$

y sea pequeño el residuo inicial $x_0 - \varphi(x_0)$ de modo que

$$|x_0 - \varphi(x_0)| \leqslant (1 - q)\,\delta. \tag{8}$$

En este caso son justas las afirmaciones:

— todas las iteraciones x_n $(n = 1, 2, \ldots)$ pertenecen al intervalo $\Delta: x_n \in \Delta$;

— la sucesión $\{x_n\}$ converge, para $n \to \infty$, hacia el límite x^* que es la raíz de la ecuación (8);

— la ecuación (2) tiene en Δ una sola raíz.

La condición $x_k \in \Delta$ significa que

$$|x_k - x_0| < \delta. \tag{9}$$

En virtud de (8) tenemos $|x_1 - x_0| = |\varphi(x_0) - x_0| \leqslant \leqslant (1 - q)\,\delta < \delta$, es decir, (9) se cumple para $k = 1$. Demostremos por el método de inducción que (9) se verifica para cualesquiera $k = 1, 2, \ldots$. Supongamos que (9) se verifica para $k = 1, 2, \ldots, n$; entonces se pueden calcular $\varphi(x_n)$ y $x_{n+1} = \varphi(x_n)$. De (6) se deduce que $|x_{k+1} - x_k| = = |\varphi(x_k) - \varphi(x_{k-1})| \leqslant q\,|x_k - x_{k-1}|$, es decir,

$$|x_{k-1} - x_k| \leqslant q\,|x_k - x_{k-1}|. \tag{10}$$

Aplicando sucesivamente esta desigualdad, encontramos

$$|x_{k+1} - x_k| \leqslant q^k\,|x_1 - x_0|, \qquad k = 1, 2, \ldots, n. \tag{11}$$

Al tomar en consideración que $x_{n+1} - x_0 = (x_{n+1} - x_n) + (x_n - x_{n-1}) + \ldots + (x_2 - x_1) + (x_1 - x_0)$, obtendremos

$$| x_{n+1} - x_0 \leqslant (q^n + q^{n-1} + \ldots + q + 1) \mid x_1 - x_0 \mid =$$
$$= \frac{1-q^{n+1}}{1-q} \mid x_1 - x_0 \mid < \frac{1}{1-q} \mid x_1 - x_0 \mid < \delta,$$

es decir, $x_{n+1} \in \Delta$. En virtud de (8), la desigualdad (9) se verifica para $k = 1$, y, por lo tanto, se verifica también para $k = 2, 3, \ldots$

Veamos ahora la diferencia $x_{n+m} - x_n = (x_{n+m} - x_{n+m-1}) + (x_{n+m-1} - x_{n+m+2}) + \ldots + (x_{n+2} - x_{n+1}) + (x_{n+1} - x_n)$ y estimémosla:

$$| x_{n+m} - x_n | \leqslant (q^{m-1} + q^{m-2} + \ldots + q + 1) \mid x_{n+1} - x_n \mid \leqslant$$
$$\leqslant \frac{1-q^m}{1-q} q^n \mid x_1 - x_0 \mid < q^n \delta,$$

es decir, $| x_{n+m} - x_n | \to 0$ para $n \to \infty$ y cualquier $m = 1, 2, \ldots$. De aquí, en virtud del criterio de Cauchy, proviene la convergencia de $\{x_n\}$: $\lim\limits_{n \to \infty} x_n = x^* \in \Delta$. Pasando ahora en (3) al límite para $n \to \infty$, nos convencemos de que x^* es una raíz de la ecuación (2): $x^* = \varphi(x^*)$. Esta raíz es única. En efecto, supongamos que existen dos raíces distintas x' y $x'' \neq x'$, de modo que $x' = \varphi(x')$, $x'' = \varphi(x'')$. Entonces. $| x'' - x' | = | \varphi(x'') - \varphi(x') \leqslant q \mid x'' - x' \mid < \mid x'' - x' \mid$, es decir, $| x'' - x' | \leqslant |x'' - x'|$, lo que no es posible.

Para el error $z_{n+1} = x_{n+1} - x^*$ tenemos

$$| z_{n+1} | = | \varphi(x_n) - \varphi(x^*) | \leqslant q \mid x_n - x^* \mid =$$
$$= q \mid z_n \mid \leqslant q^{n+1} \mid z_0 \mid, \qquad (12)$$
$$| z_{n+1} | \leqslant q^{n+1} \mid z_0 \mid,$$

es decir, el método de iteración simple converge con la velocidad de una progresión geométrica. El número de iteraciones para el cual queda cumplida la desigualdad (4) se determina de la condición $q^n \leqslant \varepsilon$, es decir,

$$n \geqslant \ln \frac{1}{\varepsilon} \Big/ \ln \frac{1}{q}.$$

El número mínimo de iteraciones $n_0(\varepsilon)$ para las cuales se cumple (4) es, evidentemente, igual a

$$n_0(\varepsilon)=\left[\ln\frac{1}{\varepsilon}\Big/\ln\frac{1}{q}\right],$$

donde $[a]$ es la parte entera del número $a > 0$.

OBSERVACIONES. Si $\varphi(x)$ tiene derivada en Δ, entonces (6) se cumple en el caso en que

$$|\varphi'(x)| \leqslant q \text{ para todo } x \in \Delta. \qquad (14)$$

2. Método de Newton. El método se determina mediante la fórmula

$$x_{n+1}=x_n-\frac{f(x_n)}{f'(x_n)}, \quad f'(x_n)\neq 0, \quad n=0,\ 1,\ 2,\ \ldots \qquad (15)$$

Esta fórmula se obtiene, si en el desarrollo

$$0=f(x^*)=f(x_n)+(x^*-x_n)f'(x_n)+\frac{1}{2}(x^*-x_n)^2f''(\xi),$$
$$\xi=x_n+\theta(x^*-x_n) \qquad 0\leqslant\theta\leqslant 1, \qquad (16)$$

donde x^* es la solución exacta de la ecuación $f(x) = 0$, se desecha el último término, al sustituir x^* por x_{n+1}:

$$0 = f(x_n) + f'(x_n)(x_{n+1} - x_n).$$

El método de Newton se denomina también *método de tangentes o de linearización*. La interpretación geométrica de este método consiste en que un trozo de la curva $y = f(x)$ para $x \in [x_n, x_{n+1}]$, si $x_n < x_{n+1}$ (o bien para $x \in [x_{n+1}, x_n]$, si $x_n > x_{n+1}$) se sustituye por el segmento de una tangente trazada desde punto $x = x_n$.

Al escribir $f(x) = 0$ en la forma $x = \varphi(x)$, vemos que el método de Newton puede ser considerado como el método de iteración simple (3) con el segundo miembro

$$\varphi(x) = x - f(x)/f'(x). \qquad (17)$$

Ilustremos el método de Newton con el ejemplo de extracción de una raíz cuadrada de un número $a > 0$, es decir, de resolución de la ecuación $x^2 = a$ o $f(x) = x^2 - a = 0$. Al aplicar la fórmula (15), obtendremos

$$x_{n+1}=\frac{1}{2}\left(x_n+\frac{a}{x_n}\right), \qquad n=0,\ 1,\ \ldots$$

Sea, $a = 2$. Eligiendo $x_0 = 1$, hallemos $x_1 = 1{,}5$, $x_2 = 1{,}417$, $x_3 = 1{,}414$, . . ., es decir, la iteración converge muy rápidamente.

Estimemos la velocidad de convergencia de las iteraciones. Supongamos que existe una raíz real x^* de la ecuación (1). Tomemos cierto entorno de la raíz:

$$\Delta_0 = (x^* - \delta_0,\ x^* + \delta_0), \qquad \delta_0 > 0.$$

Convengamos en considerar que la función (17) es dos veces diferenciable en Δ_0 y que su derivada segunda *está acotada*

$$|\varphi''(x)| \leqslant 2q, \tag{18}$$

donde $q > 0$ es una constante. Desarrollemos $\varphi(x)$ en una línea de Taylor en el entorno de $x = x^*$:

$$\varphi(x) = \varphi(x^*) + \varphi'(x^*)(x - x^*) + \frac{\varphi''(\xi)}{2}(x - x^*)^2,$$
$$\xi = x^* + \theta(x - x^*), \qquad 0 \leqslant \theta \leqslant 1. \tag{19}$$

Calculando a continuación

$$\varphi'(x) = ff''(f')^2 = -f(1/f')', \quad \psi''(x) = -\left(f\left(\frac{1}{f'}\right)'\right)'$$

y observando que $\varphi'(x^*) = 0$ cuando $f'(x^*) \neq 0$, obtendremos

$$\varphi(x_n) = \varphi(x^*) + \frac{(x_n - x^*)^2}{2}\varphi''(\xi). \tag{20}$$

Para el error $z_{n+1} = x_{n+1} - x^*$ obtendremos una fórmula:

$$z_{n+1} = x_{n+1} - x^* = \varphi(x_n) - \varphi(x^*) = \frac{1}{2}(x_n - x^*)^2\varphi''(\xi),$$
$$z_{n+1} = \frac{1}{2}\varphi''(\xi)z_n^2.$$

De aquí y de (20) se desprende

$$|z_{n+1}| \leqslant qz_n^2. \tag{21}$$

Denotando $v_n = q|z_n|$, obtenemos $v_{n+1} \leqslant v_n^2 \leqslant v_{n-1}^{22} \leqslant \ldots \leqslant$ $\leqslant \ldots \leqslant v_1^{2^n} \leqslant v_0^{2^{n+1}}$, y, *por consiguiente*,

$$|z_{n+1}| \leqslant \frac{1}{q}(q|z_0|)^{2^{n+1}}. \tag{22}$$

De aquí se ve que las iteraciones (15) convergen hacia la raíz x^* para $n \to \infty$, si

$$q \mid z_0 \mid < 1 \text{ o } \mid z_0 \mid = \mid x_0 - x^* \mid < 1/q, \qquad (23)$$

es decir, la aproximación inicial se dispone en el entorno $\Delta_0 = (x^* - 1/q,\ x^* + 1/q)$ con $\delta_0 = 1/q$ de la raíz $x = x^*$ de la ecuación (1). En este caso el método de Newton converge, como suele decirse, con la *velocidad cuadrática* (el método de iteración simple converge con la velocidad de una progresión geométrica).

La condición para que terminen las iteraciones $\mid z_n \mid \leqslant \varepsilon \, |z_0|$ (como se infiere de (22))) o $\mid z_n \mid \leqslant (q \mid z_0 \mid)^{2^n-1} \times \mid z_0 \mid$ se cumple, si $n \geqslant n_0(\varepsilon)$, donde

$$n_0(\varepsilon) = \left[\ln \left(1 + \ln \frac{1}{\varepsilon} \Big/ \ln \frac{1}{q \mid z_0 \mid} \right) \Big/ \ln 2 \right]. \qquad (24)$$

Es evidente que si la aproximación inicial se dipone en el entorno pequeño de x^*, entonces todas las iteraciones posteriores quedarán dentro de este entorno Δ_0. En efecto, sea $\mid x_0 - x^* \mid \leqslant \delta_0$, con la particularidad de que $q\delta_0 < 1$. Tendremos, pues, $\mid x_1 - x^* \mid \leqslant q \mid x_0 - x^* \mid^2 \; q\delta_0^2 < \delta_0$, $\mid x_2 - x^* \mid \leqslant q \mid x_1 - x^* \mid^2 \leqslant q\delta_0 < \delta_0$, etc., de suerte que $\mid x_n - x^* \mid \leqslant \delta_0$ para cualquier $n = 1, 2, \ldots$

OBSERVACIONES. 1. No nos detenemos en la demostración de la existencia de la raíz $x = x^*$.

2. La convergencia cuadrática del método de Newton puede establecerse también para las restricciones más débiles impuestas sobre $f(x)$:

$$|f'(x)| \geqslant M_1 > 0, \quad |f''(x)| \leqslant M_2 \text{ para todo } x \in \Delta_0. \qquad (25)$$

Haciendo uso de (15) y (16), obtendremos para el error $z_{n+1} = x_{n+1} - x^*$ una expresión

$$z_{n+1} = \frac{f''(\xi)}{2f'(x_n)} z_n^2,$$

de la cual, en virtud de las condiciones (25), proviene una desigualdad

$$|z_{n+1}| \leqslant q|z_n|^2, \qquad q = M_2/(2M_1),$$

que coincide con (21) (la diferencia consiste sólo en q). Los razonamientos ulteriores nos llevan a (22), (23), y (24).

3. Método continuo de Newton. La solución de la ecuación $f(x) = 0$ puede considerarse como un límite, para $t \to \infty$, de la solución del problema de Cauchy:

$$\frac{dx}{dt} + f(x) = 0, \qquad x > 0, \qquad x(0) = u_0, \qquad (26)$$

si este límite existe. Denotemos con $x = x(t)$ la solución del problema de Cauchy, y con x_*, la solución de la ecuación $f(x) = 0$. Para su diferencia $z(t) = x(t) - x_*$, tenemos

$$\frac{dz}{dt} + (f(x) - f(x_*)) = \frac{dz}{dt} + f'(\xi) \cdot z, \quad \xi = x_* + \theta z, \quad 0 \leqslant \theta \leqslant 1,$$

$$\frac{dz}{dt} + \alpha(t) z = 0, \qquad t > 0, \qquad z(0) = u_0, \qquad \alpha(t) = f'(\xi).$$

De aquí se ve que $|z(t)| \to 0$ para $t \to \infty$, si $f'(x) > 0$.

Para resolver la ecuación (26) se debe hacer uso de un método explícito cualquiera. La velocidad de convergencia de $x(t)$ a x_0 depende sólo de la magnitud de la derivada $f(x)$.

4. Método de las secantes. El cálculo de la derivada $f'(x_n)$, aplicado el método de Newton, puede resultar engorroso. Si sustituimos f'_n por una razón de diferencias $(f_n - f_{n-1})/(x_n - x_{n-1})$, obtendremos el método iterativo de las secantes

$$x_{n+1} = x_n - \frac{(x_n - x_{n-1}) f(x_n)}{f(x_n) - f(x_{n-1})}. \qquad (27)$$

El método de las secantes converge con una velocidad menor que el de Newton, sin embargo en (27) se calcula sólo la función, mientras que en (15) es necesario hallar no sólo la función sino también la derivada de ella. Es por esto que el volumen de los cálculos en cada iteración del método de las secantes es, en el caso general, menor.

Capítulo IV

Métodos de diferencias de la resolución de los problemas de contorno para ecuaciones diferenciales ordinarias

§ 1. Conceptos fundamentales de la teoría de esquemas de diferencias

Un método numérico universal para resolver ecuaciones diferenciales es el de diferencias finitas. Antes de pasar a su exposición hace falta introducir ciertos conceptos fundamentales referentes a la teoría de esquemas de diferencias, a saber, aproximación, estabilidad y convergencia.

1. Operadores de diferencias más simples. Con el fin de obtener una ecuación en diferencias en lugar de la ecuación diferencial, es necesario:

— sustituir el dominio de variación continua del argumento por un conjunto discreto de puntos (por una red);

— sustituir (aproximar en la red) la ecuación diferencial por una ecuación en diferencias.

El problema sobre la resolución numérica de una ecuación diferencial se reduce a la cuestión de resolver las ecuaciones en diferencias. En los capítulos antecedentes ya se han expuesto los ejemplos de las redes:

1) red uniforme en el segmento $0 \leqslant x \leqslant 1$ de paso h: un conjunto de nodos $\bar{\omega}_h = \{x_i = ih,\ i = 0,\ 1,\ 2,\ \ldots,\ N,\ h = 1/N\}$; $x_0 = 0$, $x_N = 1$ son los nodos de frontera; $\omega_h = \{x_i = ih,\ i = 1,\ 2,\ \ldots,\ N\text{-}1\}$ es el conjunto de nodos interiores;

2) red no uniforme: el segmento $0 \leqslant x \leqslant 1$ se divide en N partes mediante puntos arbitrarios $x_1 < x_2 < \ldots < x_{N-1}$; $h_i = x_i - x_{i-1}$ es el paso de la red;

$$\bar{\omega}_h = \{x_i,\ i = 0,\ 1,\ \ldots,\ N,\ x_0 = 0,\ x_N = 1\},$$

$$\sum_{i=1}^{N} h_i = 1 \qquad \omega_h = \{x_i,\ 0 < i < N\};$$

3) red en un segmento $0 \leqslant t \leqslant T$: $\bar{\omega} = \{t_n = n\tau, n = 0, 1, \ldots, n_0; n_0\tau = T\}$.

En lugar de la función de argumento continuo (por ejemplo, en el segmento $0 \leqslant x \leqslant 1$) se estudia la función $y(x_i) = y_i$ de argumento discreto x_i, donde x_i es un nodo de la red $\bar{\omega}_h$, o de argumento i que es el número del nodo. Esta función se denomina *reticular*. Cualquier función reticular puede ser representada en forma de un vector

$$Y = y_0, y_1, \ldots, y_{N-1}, y_N).$$

Por eso, el conjunto de funciones reticulares forma un espacio de dimensión finita H cuya dimensión, en el caso dado, es $(N + 1)$. Se analiza corrientemente una familia de redes $\{\omega_h\}$ que dependen del paso como de un parámetro, razón por la cual las funciones reticulares $y = y_h(x)$ también dependen del paso como de un parámetro (o de N), si la red ω_h es uniforme. Si la red no es uniforme, entonces por h se entiende un vector $h = (h_1, h_2, \ldots, h_N)$. Resulta natural por esta razón dotar el espacio de funciones reticulares con el índice h y escribir H_h. En el espacio H_h podemos introducir una norma $\|\cdot\|_h$. He qauí los tipos más simples de las normas:

$$\|y\|_C = \max_{x \in \bar{\omega}_h} y(x) \text{ o bien } \|y\|_C = \max_{0 \leqslant i \leqslant N} |y_i|;$$

$$\|y\| = \left(\sum_{i=1}^{N-1} y_i^2 h\right)^{1/2}.$$

El operador diferencial se sustituye por un operador de diferencias que actúa en el espacio de funciones reticulares.

Sea G un dominio del espacio euclídeo R^p $(p = 1, 2, 3)$ con la frontera Γ. Por ejemplo, G es el intervalo $0 < x < 1$, Γ expresa los puntos $x = 0$, $x = 1$; G es un rectángulo $0 < x_1 < l_1$, $0 < x_2 < l_2$, $x = (x_1, x_2) \in G$ $(p = 2)$, Γ consta de los segmentos de las rectas $x_2 = 0$, $x_2 = l_2$, $x_1 = 0$, $x_1 = l_1$, etc. Sea dado un operador diferencial lineal L que actúa contra una función $v(x)$, $x \in G$. Introduzcamos en $G = \bar{G} \cup \Gamma$ una red $\bar{\omega}_h$ y veamos una función reticular $v_h(x)$, $x \in \omega_h$. Sustituyamos Lv en el punto $x_i \in \omega_h$ por una combinación lineal consistente de los valores $v_h(x)$ de la función reticular en cierto conjunto de nodos de la red

el cual se llamará *molde*

$$(L_h v)_i = \sum_{x_j \in \sigma_i} a_{ij}^h v_h(x_j), \quad x_i \in \omega_h(G), \tag{1}$$

donde a_{ij}^h son los coeficientes, σ_i es el molde, $\sigma_i \in \overline{\omega}_h$.

Tal sustitución de Lv por $L_h v$ se denomina *aproximación en la red* de un operador diferencial L mediante el operador de diferencias L_h, o bien *aproximación de diferencias* del operador L. El estudio de las aproximaciones de diferencias L_h del operador L se realiza, corrientemente, de un modo local, es decir, en cualquier punto fijo de la red. La construcción de L_h se debe empezar con la elección de un molde $\sigma(x)$, es decir, de un conjunto de nodos, vecinos con el nodo $x \in \omega_h$, en los cuales los valores de la función reticular $v_h(x)$ pueden ser empleados al escribir la expresión para L_h.

Veamos algunos ejemplos de construcción de L_h.

EJEMPLO 1. Derivada primera: $Lv = \frac{dv}{dx} = v'(x)$. Tomemos tres nodos $(x-h,\ x,\ x+h)$. Podemos servirnos de cualquier de las expresiones

$$L_h^+ v = \frac{v(x+h) - v(x)}{h} = v_x \quad \text{(el molde } (x,\ x+h));$$

$$L_h^- v = \frac{v(x) - v(x-h)}{h} = v_{\bar{x}} \quad \text{(el molde } (x-h,\ x));$$

$$L_h^0 v = \frac{v(x+h) - v(x-h)}{2h} = v_{\overset{\circ}{x}} \quad \text{(el molde } (x-h,\ x+h).$$

Se emplean a menudo las siguientes denominaciones: $L_h^+ v = v_x$ es la derivada de diferencias *derecha*; $L_h^- v = v_{\bar{x}}$, la derivada de diferencias *izquierda* y $L_h^0 v = v_{\overset{\circ}{x}} = \frac{1}{2}(L_h^+ v + L_h^- v)$, *derivada de diferencias central*. Sobre el molde tripuntual $(x-h,\ x,\ x+h)$ podemos definir un operador de diferencias

$$L_h^{(\sigma)}(v) = \sigma v_x + (1-\sigma) v_{\bar{x}},$$

donde σ es un parámetro real. De este modo, existe una infinidad de aproximaciones de diferencias de la primera derivada sobre el molde tripuntual.

Se denomina *error de aproximación del operador L* mediante el operador L_h una diferencia

$$\psi = L_h v - Lv.$$

Dicen que L_h *tiene el m-ésimo orden de aproximación en el punto x*, si

$$\psi(x) = L_h v(x) - Lv(x) = O(h^m), \text{ o bien } |\psi(x)| \leqslant Mh^m,$$

donde $M = \text{const} > 0$ no depende de h, $m > 0$.

Haciendo uso de la fórmula de Taylor

$$v(x \pm h) = v(x) \pm hv'(x) + \\ + \frac{h^2}{2} v''(x) \pm \frac{h^3}{6} v'''(x) + \frac{h^4}{24} v^{IV}(x) + O(h^5),$$

no será difícil obtener las estimaciones

$$v_x - v' = O(h), \quad v_{\bar{x}} - v' = O(h), \quad v_{\overset{\circ}{x}} - v' = O(h^2),$$

$$\psi^{(\sigma)} = L_h^{(\sigma)} v - Lv = O\left(\left(\sigma - \frac{1}{2}\right) h + h^2.\right.$$

EJEMPLO 2. Derivada segunda: $Lv = \frac{d^2v}{dx^2} = v''(x)$.

Tomemos el mismo molde tripuntual que figuraba en el ejemplo 1 y escribamos un operador de diferencias

$$L_h v(x) = \frac{v(x+h) - 2v(x) + v(x-h)}{h^2}.$$

Al notar que $v(x+h) = v(x) + hx_x$, $v(x-h) = v(x) - hv_{\bar{x}}$, transformemos $L_h\ v(x)$:

$$L_h v(x) = \frac{v_x(x) - v_{\bar{x}}(x)}{h} = \frac{v_{\bar{x}}(x+h) - v_{\bar{x}}(x)}{h} = v_{\bar{x}x}(x). \quad (2)$$

Aprovechando la fórmula de Taylor para $v(x \pm h)$, encontramos

$$\psi = L_h v - Lv = \frac{h^2}{12} v^{IV}(x) + O(h^4) = O(h^2),$$

es decir, L_h tiene el segundo orden de aproximación.

Habitualmente se requiere la estimación del error de aproximación sobre una red, es decir, en cierta norma reticular $\|\cdot\|_h$. Se dice que L_h tiene el m-ésimo *orden de aproxima-*

ción sobre una red, siempre que

$$\|L_h v_h - (Lv)_h\|_h = O(h^m).$$

2. **Esquema de diferencias.** La ecuación diferencial $Lu = f(x)$ se resuelve, como regla, con ciertas condiciones complementarias: iniciales (problemas de Cauchy), de contorno (problemas de contorno) o bien condiciones iniciales y las de contorno a la vez. Dichas condiciones complementarias, pasando a las ecuaciones en diferencias, se deben también aproximar.

Sea dado un dominio G con la frontera Γ y supongamos que se busca la solución $u = u(x)$, $x \in \bar{G}$, de una ecuación diferencial lineal

$$Lu = f(x), \qquad x \in G, \tag{3}$$

con la siguiente condición complementaria en la frontera:

$$u(x) = \mu(x), \qquad x \in \Gamma. \tag{4}$$

Introduzcamos en el dominio $\bar{G} = G + \Gamma$ una red $\bar{\omega}_h = \omega_h + \gamma_h$, $\omega_h \in G$, $\gamma_h \in \Gamma$, y al problema (3), (4) le pondremos en correspondencia un problema de diferencias con el operador lineal L_h del tipo (1):

$$L_h y_h = \varphi_h(x), \quad x \in \omega_h; \quad y_h(x) = \nu_h(x), \quad x \in \gamma_h. \tag{5}$$

Las funciones $y_h(x)$, $\varphi_h(x)$, $\nu_h(x)$ dependen del paso h de la red. Al variar h, obtenemos las sucesiones $\{y_h\}$, $\{\varphi_h\}$, $\{\nu_h\}$. De este modo, se examina no uno de los problemas de diferencias, sino una familia de problemas que depende del parámetro h. Esta familia de problemas lleva el nombre de *esquema de diferencias*.

EJEMPLO 1. Problema de Cauchy:

$$Lu = \frac{du}{dt} + \lambda u = f(t), \quad t > 0, \quad u(0) = u_0.$$

El esquema de diferencias de Euler tiene por expresión:

$$L_\tau y = \frac{y_{n+1} - y_n}{\tau} + \lambda y_n = f_n,$$

$$y_n = y(t_n), \quad t_n = n\tau \in \omega_\tau, \quad n = 0, 1, \ldots, y_0 = u_0.$$

EJEMPLO 2. Primer problema de contorno:

$$Lu = u'' = -f(x), \quad 0 < x < 1, \quad u(0) = \mu_1, \quad u(1) = \mu_2. \qquad (6)$$

Hagamos uso del operador de diferencias tripuntual (2): $L_h y_i = y_{\bar{x}x,i} = (y_{i+1} - 2y_i + y_{i-1})/h^2$ y obtendremos un problema de contorno en diferencias sobre la red $\bar{\omega}_h = \{x_i = ih,\ 0 \leqslant i \leqslant N,\ x_N = 1\}$:

$$L_h y_i = y_{\bar{x}x,i} = -f_i, \quad i = 1, 2, \ldots, N-1, \quad y_0 = \mu_1,\ y_N = \mu_2. \qquad (6')$$

3. Estabilidad. Nos resulta más conveniente pasar a la notación del esquema de diferencias (5) en la forma operacional. Con este fin escribamos al principio la ecuación (5) en la forma matricial

$$AY_h = \Phi_h,$$

donde Y_h es el vector buscado de N-ésima dimensión finita, la que es igual al número de nodos de la red, en los cuales no son conocidos los valores de la función reticular y_h (para el primer problema de contorno (6′) la dimensión Y_h es igual a N-1, es decir, al número de nodos interiores de la red). Los valores de $y_h(x_i)$ en los nodos $x_i \in \omega_h$ son componentes del vector Y_h, mientras que $\varphi_h(x_i)$ representan los componentes del vector Φ_h, y A es la matriz cuadrada de dimensión $N \times N$.

Introduzcamos un espacio N-dimensional H_h de funciones reticulares y sea A_h un operador lineal correspondiente a la matriz A: $A_h : H_h \to H_h$. En lugar de (7) podemos escribir

$$A_h y_h = \varphi_h, \qquad \varphi_h \in H_h. \qquad (8)$$

Sean $\|\cdot\|_{(1_h)}$ y $\|\cdot\|_{(2_h)}$ ciertas normas en el espacio H_h.

Diremos que el esquema de diferencias (8) es estable, si existe una constante $M > 0$ (y dicha constante no depende de h ni del modo de elegir φ_h) tal que para la solución y_h de la ecuación (8) tiene lugar la estimación

$$\|y_h\|_{(1_h)} \leqslant M\, \|\varphi_h\|_{(2_h)} \qquad (9)$$

con todo h suficientemente pequeño: $|h| \leqslant h_0$.

El esquema de diferencias (8) se denomina *correcto* (*correctamente planteado*), si la solución de la ecuación (8) existe y es única, cualesquiera que sean los datos de entrada de $\varphi_h \in H_h$, y si el esquema de diferencias es estable, es decir, queda cumplida la desigualdad (9).

La estabilidad del esquema significa una dependencia continua de la solución y_h de los datos de entrada, con la particularidad de que dicha dependencia continua es uniforme respecto de h. Si $\tilde{y}_h$ es una solución de la ecuación $A_h\tilde{y}_h = \tilde{\varphi}_h$, entonces $A_h(\tilde{y}_h - y_h) = \tilde{\varphi}_h - \varphi_h$ en virtud de la linealidad de A_h; en este caso, de (9) proviene

$$\| \tilde{y}_h - y_h \|_{(1_h)} \leqslant M \| \tilde{\varphi}_h - \varphi_h \|_{(2_h)}. \qquad (10)$$

A la variación pequeña de los datos de entrada le corresponde la variación pequeña de la solución.

Si el esquema (8) es resoluble, existe un operador inverso A_h^{-1} y

$$y_h = A_h^{-1}\varphi_h, \quad \| y_h \|_{(1_h)} \leqslant \| A_h^{-1} \| \, \| \varphi_h \|_{(2_h)}, \qquad (11)$$

donde $\| A_h^{-1} \| = \| A_h^{-1} \|_{(2_h \to 1_h)}$ es la norma del operador A_h^{-1}.

La estabilidad es un testimonio de que el operador inverso está acotado uniformemente respecto de h

$$\| A_h^{-1} \| \leqslant M. \qquad (12)$$

El esquema es *inestable* si no existe tal constante M (*no dependiente de h*) *que supere* $\|A_h^{-1}\|$, *es decir*, $\|A_h^{-1}\|$ crece indefinidamente cuando $|h| \to 0$.

Puede suceder que en lugar de la condición de contorno de la primera especie $u = \mu$ para $x \in \Gamma$ viene prefijada la condición

$$lu = \mu(x), \qquad x \in \Gamma, \qquad (13)$$

donde l es cierto operador diferencial lineal, por ejemplo, $lu = u' - \sigma u$, $\sigma > 0$, o bien $lu = u'$ para $x = 0$ ó para $x = 1$. Entonces, en vez del problema (3), (4) tenemos el siguiente

$$Lu = f(x), \quad x \in G; \quad lu = \mu(x), \quad x \in \Gamma. \qquad (14)$$

El esquema de diferencias correspondiente tendrá

$$L_h y_h = \varphi_h \text{ para } x \in \omega_h, \quad l_h y_h = \bar{\mu}_h \quad \text{para } x \in \gamma_h, \tag{15}$$

donde l_h es un operador de diferencias lineal que aproxima el operador l. Puede ocurrir, además, que φ_h y $\bar{\mu}_h$ han de ser estimadas en las normas diferentes $\|\varphi_h\|_{(2_h)}$, $\|\bar{\mu}_h\|_{(3_h)}$.

El esquema (15) es *estable* si para su solución y_h queda válida la estimación

$$\|y_h\|_{(1_h)} \leqslant M_1 \|\varphi_h\|_{(2_h)} + M_2 \|\mu_h\|_{(3_h)}, \tag{16}$$

donde $M_1 > 0$, $M_2 > 0$ son unas constantes que no dependen ni de h ni del modo de elegir los datos de entrade φ_h y $\bar{\mu}_h$.

Se ha de notar que el esquema de diferencias (15) también puede ser escrito en la forma operacional $A_h y_h = \varphi_h$, sin embargo, en este caso, $\|\cdot\|_{(2_h)}$ en (9) y (16) pueden diferir, al igual que los propios miembros segundos (lo que ya está claro para el primer problema de contorno).

4. Ejemplo de un esquema estable. A título de ejemplo de un esquema estable analicemos el siguiente problema de contorno en diferencias

$$y_{\bar{x}x,\,i} = \frac{y_{i-1} - 2y_i + y_{i+1}}{h^2} = -\varphi_i, \quad i = 1, 2, \ldots, N-1,$$
$$y_0 = 0, \qquad y_N = 0, \qquad hN = 1. \tag{17}$$

Siguiendo las indicaciones del § 4, cap. I, definamos el operador A_h. Sea H_h un espacio de funciones reticulares definidas en los nodos interiores ($i = 1, 2, \ldots, N-1$) de la red. Tomemos $y \in H_h$ (el índice h de $y_h(x)$ queda por ahora omitido) y una función $\dot{y}_0 = \dot{y}_N = 0$. Entonces, el operador A_h se determina con ayuda de la identidad

$$(A_h y)_i = -\dot{y}_{\bar{x}x,\,i}, \qquad i = 1, 2, \ldots, N-1,$$

y en lugar de (17) se obtiene una ecuación operacional

$$A_h y_h = \varphi_h. \tag{18}$$

En el espacio H_h introducimos el producto escalar

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

El operador A_h en H_h es autoconjugado y definido positivo y

$$\delta E \leqslant A_h \leqslant \Delta E, \text{ o bien } \delta \| y \|^2 \leqslant (A_h y, y) \leqslant \Delta \| y \|^2 \quad \text{para todo } y \in H_h, \tag{19}$$

donde δ y Δ son los valores propios mínimo y máximo, respectivamente, del operador A:

$$\delta = \frac{4}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}, \qquad \Delta = \| A_h \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}. \tag{20}$$

El operador inverso A_h^{-1} es autoconjugado si $A_h = A_h^*$. En el § 4, cap I, se ha mostrado que las desigualdades (19) son equivalentes a las desigualdades operacionales

$$\frac{1}{\Delta} E \leqslant A_h^{-1} \leqslant \frac{1}{\delta} E, \quad \| A_h^{-1} \| = \frac{1}{\delta}. \tag{21}$$

De aquí se deducen la acotación uniforme de la norma del operador inverso A_h^{-1} : $\|A_h^{-1}\| \leqslant 1/\delta < 1/8$ y la estimación apriorística

$$\| y_h \| \leqslant \frac{1}{\delta} \| \varphi_h \| \leqslant \frac{1}{8} \| \varphi_h \|, \tag{22}$$

que es indicio de estabilidad del esquema (18). Esta estimación puede obtenerse por el método de desigualdades energéticas, sin recurrir a la estimación de los valores propios $\lambda_k (A_h^{-1})$. En efecto, multipliquemos la ecuación $A_h y_h = \varphi_h$ escalarmente por y_h: $(A_h y_h, y_h) = \varphi_h, y_h)$ y aprovechemos las desigualdades $(\varphi_h, y_h) \leqslant \| \varphi_h \| \| y_h \|$, $\| y_h \|^2 \leqslant \leqslant \frac{1}{\delta} (A_h y_h, y_h)$; entonces obtendremos la desigualdad $\delta \| y_h \|^2 \leqslant \| \varphi_h \| \| y_h \|$, de donde se deduce precisamente la estimación (22).

El esquema (17) es estable también en la norma $\| y \|_C$:

$$\| y_h \|_C \leqslant \frac{1}{2} \| \varphi_h \|_C, \quad \| y \|_C = \| y \|_{C_h} = \max_{0<i<N} |y_i|. \tag{23}$$

Esto proviene de la estimación de la solución del problema de controrno en diferencia tripuntual, obtenido en el p. 3 del § 5, cap. I. En el caso dado la estimación tiene por

expresión

$$\| y_h \|_C \leqslant \sum_{s=1}^{N-1} h \sum_{k=1}^{s} h |\varphi_k| \leqslant \| \varphi \|_C \sum_{s=1}^{N} x_s h < \frac{1}{2} \| \varphi_h \|_C,$$

puesto que

$$\sum_{s=1}^{N} x_s h = h^2 \sum_{s=1}^{N} s = \frac{N(N-1)}{2} h_2 = \frac{1-h}{2} < \frac{1}{2}.$$

5. Ejemplo de un esquema no correcto. Sea dado un esquema

$$A_h y_h = \varphi_h$$

y $\| A_h \| \to \infty$ cuando $|h| \to 0$. Veamos un problema inverso: determinar el segundo miembro φ_h por la solución conocida y_h:

$$B_h \varphi_h = y_h, \qquad B_h = A_h^{-1}.$$

El problema no es correctamente planteado, puesto que

$$\| B_h^{-1} \| = (A_h^{-1})^{-1} \| = \| A_h \| \to \infty \text{ cuando } |h| \to 0.$$

Esto significa que para cualquier constante M que no depende de h puede indicarse tal h_* que $\| B_h^{-1} \| > M$ cuando $|h| \leqslant |h_*|$. Sea $\tilde{\varphi}_h$ la solución de la ecuación $B_h \varphi_h = \tilde{y}_h$, y sea φ_h la solución de la ecuación $B_h \varphi_h = y_h$, entonces

$$\| \tilde{\varphi}_h - \varphi_h \| \leqslant \| B_h^{-1} \| \| \tilde{y}_h - y_h \|.$$

Si, en cambio,

$$\| B_h^{-1} \| \leqslant M \| \text{ para } |h| \geqslant h_0,$$

de modo que se verifica la desigualdad

$$\| \tilde{\varphi}_h - \varphi_h \| \leqslant M \| \tilde{y}_h - y_h \|,$$

diremos que el esquema es *casi estable*. ¿Se podrá aplicar este esquema para determinar φ_h con la exactitud requerida ε, si y_h viene prefijada con cierta exactitud ε_0:

$$\| \tilde{y}_h - y_h \| \leqslant \varepsilon?$$

De la desigualdad $\| \tilde{\varphi}_h - \varphi_h \| \leqslant \| B_h^{-1} \| \; \| \tilde{y}_h - y_h \|$ se desprende que la solución del problema $B_h \varphi_h = y_h$ se determina con la exactitud $\| B_h^{-1} \| \varepsilon_0$. Supongamos que se pide hallar φ_h con la exactitud $\varepsilon > 0$, de suerte que $\| \tilde{\varphi}_h - \varphi_h \| \leqslant \varepsilon$; esto es posible bajo la condición

$$\| B_h^{-1} \| \cdot \varepsilon_0 \leqslant \varepsilon.$$

De aquí determinamos el paso admisible $h \geqslant h_0$, es decir, h_0.

Expliquemos esto con el problema concreto (17). Para dicho problema tenemos

$$\| B_h^{-1} \| = \| A_h \| = \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2} \leqslant \frac{4}{h^2},$$

y la condición $\| B_h^{-1} \| \; \varepsilon_0 = \Delta \varepsilon_0 \leqslant \varepsilon$ queda cumplida si $4\varepsilon_0/h^2 \leqslant \varepsilon$, o bien

$$h \geqslant h_0 = 2\sqrt{\varepsilon_0/\varepsilon}.$$

De aquí se ve que la precisión con la que se dan los datos de entrada ε_0 ha de ser más alta que la exactitud ε con la que se determina la solución.

Por ejemplo, sean prefijados el error del segundo miembro $\varepsilon_0 = 10^{-8}$ y la exactitud requerida $\varepsilon = 10^{-4}$. Entonces, $h_0 = 2 \cdot 10^{-2} = 1/50$, es decir, la exactitud $\varepsilon = 10^{-4}$ puede obtenerse sólo sobre una red de paso $h \geqslant 1/50$. Si en cambio, por ejemplo, $\varepsilon_0 = \frac{1}{4} \cdot 10^{-4}$, $\varepsilon = 10^{-4}$, entonces $h_0 = 1$ y la exactitud $\varepsilon = 10^{-4}$ no se conseguirá en ninguna red para tal precisión de prefijar los datos de entrada.

6. Aproximación y convergenica. Al resolver el problema (14) por el método de diferencias se debe saber con qué exactitud la resolución del problema de diferencias aproxima la solución del problema inicial. Con el fin de estimar el error obtenidò al sustituir (14) por el esquema de difrencias (15), es necesario comparar las soluciones de estos problemas. La comparación se realizará en el espacio H_h de funciones reticulares. Denotemos con $u_h(x)$ los valores de las funciones $u(x)$ (soluciones exactas del problema (14)) sobre la red ω_h: $u_h \in H_h$. Veamos un error

$$z_h = y_h - u_h,$$

donde y_h es la solución del problema (15). Al sustituir $y_h = z_h + u_h$ en (15) y al tomar $u = u(x)$ por la función prefijada, obtendremos para z_h un problema de diferencias

$$L_h z_h = \psi_h, \quad x \in \omega_h; \quad l_h z_h = \nu_h, \quad x \in \gamma_h, \qquad (24)$$

donde $\psi_h = \varphi_h - L_h u_h$ se denomina *error de aproximación para la ecuación* $L_h y_h = \varphi_h$ *en la solución* $u = u(x)$ *de la ecuación* $Lu = f(x)$ *(residuo para el esquema de diferencias en la solución)*; $\nu_h = \overline{\mu_h} - l_h u_h$ recibe el nombre de error de aproximación para la condición de contorno de diferencias $l_h y_h = \overline{\mu_h}$ en la solución del problema (14).

Diremos que:

el esquema de diferencias (15) *converge*, si

$$\| z_h \|_{(1_h)} \to 0 \text{ para } |h| \to 0;$$

el esquema de diferencias (15) tiene *exactitud* de m-ésimo orden o *converge con la velocidad* $O(|h|^m)$, si

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} \leqslant M |h|^m$$

o

$$\| z_h \|_{(1_h)} = O(|h|^m), \qquad m > 0,$$

donde $M > 0$ es una constante no dependiente de h.

El esquema de diferencias (15) tiene el m-ésimo *orden de aproximación en la solución*, si

$$\| \psi_h \|_{(2h)} = O(|h|^m), \ \| \nu_h \|_{(3_h)} = O(|h|^m), \quad m > 0. \qquad (25)$$

La estimación de los residuos ψ_h y ν_h se realiza bajo el supuesto de que la solución del problema inicial existe y tiene tantas derivadas cuanto es necesario al obtener el m-ésimo orden de aproximación.

Demos a conocer dos ejemplos de estimar ψ_h.

EJEMPLOS. 1. Hay un problema

$$L_h y = -y_{\bar{x}x} = \varphi(x), \quad x = ih, \quad 1 \leqslant i \leqslant N-1,$$
$$y_0 = y_N = 0, \qquad (26)$$
$$Lu = -u'' = f(x), \quad 0 < x < 1, \quad u(0) = u(1) = 0.$$

En este caso las condiciones de contorno se satisfacen con toda la exactutud, $\nu_h = 0$ (el índice h de $\varphi(x)$, $u(x)$ por ahora queda omitido) y

$$\psi_h = \varphi - L_h u = \varphi + u_{\bar{x}x} = \varphi + \left(u'' + \frac{1}{2} h^2 u^{IV} + O(h^4)\right) =$$

$$= (\varphi + u'') + \frac{h^2}{12} u^{IV} + O(h^4) = \varphi - f + O(h^2),$$

puesto que $u'' = -f(x)$. De aquí se ve que $\|\psi_h\|_C = O(h^2)$, si ponemos $\varphi = f$, o bien $\varphi = f + O(h^2)$.

En el p. 1 fue estimado el error $\psi = L_h v_h - (Lv)_h$ para una función arbitraria. En la estimación del error $z_h = y_h - u_h$ se une el residuo ψ_h que caracteriza el error de aproximación del operador $Lu - f$ mediante el operador $L_h u_h - \varphi_h$ en la solución $u = u(x)$ del problema inicial. Al tomar en consideración que $f - Lu = 0$, representemos $\psi_h = \varphi_h - L_h u_h$ en la forma

$$\psi_h = (\varphi_h - L_h u_h) - (f - Lu)_h =$$

$$= (\varphi_h - f_h) - (L_h u_h - (Lu)_h) = \psi_h^{(1)} + \varphi_h^{(2)},$$

donde $\psi_h^{(1)} = -(L_h u_h - (Lu)_h)$, $\psi_h^{(2)} = \varphi_h - f_h$; $\psi_h^{(1)}$ es el error de aproximación de L por el operador L_h en la solución $u = u(x)$ del problema (6), $\psi_h^{(2)}$ es el error de aproximación del segundo miembro de la ecuación. La exigencia $\|\psi_h\|_{(2_h)} = O(|h|^m)$ se cumple, evidentemente, si $\|\psi_h^{(1)}\|_{(2_h)} = O(|h|^m)$, $\|\psi_h^{(2)}\|_{(2_h)} = O(|h|^m)$. No obstante, estas condiciones no son necesarias para la estimación de $\|\psi_h\|_{(2_h)} = O(|h|^m)$, lo que atestigua el siguiente ejemplo.

2. Primer problema de contorno (6). Calculemos

$$-\psi_h^{(1)} = u_{\bar{x}x} - u'' = \frac{1}{12} h^2 u^{IV} + O(h^4) = O(h^2).$$

Sea $\varphi = f + \frac{1}{12} h^2 f_{\bar{x}x}$, es decir, $\varphi - f = O(h^2)$. De aquí se ve que $\psi_h^{(1)} = O(h^2)$ y $\psi_h^{(2)} = O(h^2)$, sin embargo, el esquema tiene el cuarto orden de aproximación, puesto que

$$\psi_h = \psi_h^{(1)} + \psi_h^{(2)} = \varphi - f + \frac{h^2}{12} u^{IV} + O(h^4) =$$

$$= \frac{h^2}{12}(f_{\bar{x}x} + u^{IV}) + O(h^4) = \frac{h^2}{12}(f'' + u^{IV}) + O(h^4),$$

$\psi_h = O(h^4)$, dado que $u^{IV} + f''(x) = 0$ en virtud de la ecuación $u'' + f(x) = 0$.

7. Relación de la estabilidad y aproximación con la convergencia. Examinemos un esquema de diferencia lineal (15). Si el esquema es estable y aproxima el problema inicial, será convergente (se dice corrientemente: «de la estabilidad y aproximación proviene la convergencia del esquema»). En efecto, para el error $z_h = y_h - u_h$ obtenemos, en virtud de la linealidad de L_h y l_h, el problema (24) que es análogo al problema (15) para y_h. Por eso, si el esquema (15) es estable, es decir, si es justa la estimación (16), entonces para z_h será válida la estimación

$$\| z_h \|_{(1_h)} \leqslant M_1 \| \psi_h \|_{(2_h)} + M_2 \| \nu_h \|_{(3_h)}. \tag{27}$$

De aquí se deduce que

$$\| z_h \|_{(1_h)} = \| y_h - u_h \|_{(1_h)} = O(|h|^m).$$

siempre que

$$\| \psi_h \|_{(2_h)} = O(|h|^m), \quad \| \nu_h \|_{(3_h)} = O(|h|^m).$$

De este modo, el estudio de la convergencia y del orden de exactitud de los esquemas de diferencias se reduce al estudio del error de aproximación y de la estabilidad, es decir, a la obtención de las estimaciones apriorísticas (16).

EJEMPLO. Para el esquema de diferencias (17) ($y_{\bar{x}x,i} = -\varphi_i$, $i = 1, 2, \ldots, N-1$, $y_0 = 0$, $y_N = 0$) se ha obtenido anteriormente la estimación (23). El error de aproximación es evidentemente, $\| \psi_h \|_{C_h} = O(h^2)$ para $\varphi_i = f_i$, $\| \psi_h \|_{C_h} = O(h^4)$ para $\varphi_i = f_i + \frac{h^2}{12} f_{\bar{x}x,i}$. Por cuanto $z_{\bar{x}x,i} = -\psi_{h,i}$ para $i = 1, 2, \ldots, N-1$, $z_0 = 0$, $z_N = 0$, entonces para z será también válida la estimación $\| z \|_C \leqslant \frac{1}{2} \| \psi \|_C$, de donde se desprende $\| y_h - u_h \|_C = O(h^m)$, donde $m = 2$ para $\varphi = f$, $m = 4$ para $\varphi = f + \frac{h^2}{12} f_{\bar{x}x}$.

Con ello se da por terminado el estudio del esquema (26) (el estudio del esquema (26) se ha ilustrado, de hecho, con los tres últimos ejemplos). Todo lo expuesto más arriba sirve de ejemplo típico de cómo se realiza el estudio de los esquemas de diferencias.

§ 2. Esquemas de diferencias homogéneos tripuntuales

1. Problema de partida. Consideremos el primer problema de contorno para una ecuación diferencial ordinaria de segundo orden:

$$Lu = \frac{d}{dx}\left(k(x)\frac{du}{dx}\right) - q(x)u = -f(x), \quad 0 < x < 1,$$
$$k(x) \geqslant q > 0, \quad q(x) \geqslant 0, \; u(0) = \mu_1, \quad u(1) = \mu_2. \tag{1}$$

Una ecuación de este tipo describe la distribución estacionaria de temperatura, es decir, una distribución que no varía en tiempo (ecuación estacionaria de conductibilidad térmica), o bien la distribución de concentración (ecuación de difusión). Si $u = u(x)$ es la temperatura, entonces $W(x) = -k(x)\frac{du}{dx}$ es un flujo térmico ($k(x)$ es el coeficiente de conductibilidad térmica).

El problema (1) tiene una solución única, si $k(x)$, $q(x)$, $f(x)$ son funciones continuas a trozos. Si $k(x)$ tiene una discontinuidad de primera especie en el punto $x = \xi$, de modo que $[k] = k(\xi + 0) - k(\xi - 0) \neq 0$, en dicho punto deben ser continuos tanto la temperatura u como el flujo térmico $-(ku')$:

$$[u] = 0, \quad [ku'] = 0 \quad \text{para} \quad x = \xi.$$

Son posibles también otras condiciones de contorno para $x = 0$, $x = 1$: $ku' = \sigma_1 u - \mu_1$ para $x = 0$, $-ku' = \sigma_2 u - \mu_2$ para $x = 1$. Si $\sigma_1 > 0$, entonces la citada condición es de tercera especie; cuando $\sigma_1 = 0$, tenemos la condición de segunda especie ($ku' = -\mu_1$ para $x = 0$). Son posibles combinaciones de diferentes condiciones para $x = 0$ y $x = 1$.

2. Esquemas de diferencias tripuntuales. Introduzcamos en el segmento $0 \leqslant x \leqslant 1$ una red uniforme $\omega_h = \{x_i = ih, \; i = 0, 1, \ldots, N\}$ de paso $h = 1/N$ y elijamos un molde tripuntual (x_{i-1}, x_i, x_{i+1}), en el cual escribiremos el esquema de diferencias que aproxima el problema (1). Cualquier ecuación en diferencias en este molde tendrá por expresión

$$b_i y_{i+1} - c_i y_i + a_i y_{i-1} = -h^2 \varphi_i, \tag{2}$$

donde a_i, b_i, c_i son los coeficientes dependientes de $k(x)$, $q(x)$ y h. Estos coeficientes son, por ahora, desconocidos. Escribamos (2) de otra forma:

$$\frac{1}{h}\left(b_i\frac{y_{i+1}-y_i}{h}-a_i\frac{y_i-y_{i-1}}{h}\right)-d_iy_i=-\varphi_i, \qquad (3)$$
$$d_i=(c_i-a_i-b_i)/h^2.$$

Diremos que un esquema de diferencias es *homogéneo*, si sus coeficientes en todos los nodos de la red para cualesquiera coeficientes de la ecuación diferencial se calculan según unas mismas fórmulas. Así, por ejemplo, si introducimos las funcionales $A[\bar{k}(s)]$, $B[\bar{k}(s)]$, $D[\bar{k}(s)]$, $F(\bar{f}(s)]$, definidas para cualesquiera funciones continuas a trozos sobre el segmento $-1\leqslant s\leqslant 1$, y calculamos los coeficientes del esquema (3) por las fórmulas

$$a_i=A[k(x_i+sh)], \qquad b_i=B[k(x_i+sh)],$$
$$d_i=D[k(x_i+sh)],\ \varphi_i=F[f(x_i+sh)],\ \bar{k}(s)=k(x_i+ +sh),$$

entonces tal esquema será homogéneo. He aquí las funcionales más simples

$$A[\bar{k}(s)]=\bar{k}(-0{,}5), \qquad a_i=k_{i-1/2}=k(x_i-0{,}5h),$$
$$F(\bar{f}(s))=f(0), \qquad \varphi_i=f_i=f(x_i),\ \text{etc.}$$

Si un esquema es homogéneo, resulta más cómodo servirse del sistema de designaciones sin índices:

$$\Lambda_y=\frac{1}{h}(by_x-ay_{\bar{x}})-dy=-\varphi, \qquad x\in\omega_h,$$
$$y(0)=\mu_1, \qquad y(1)=\mu_2, \qquad (4)$$

donde

$$a=a(x), \quad b=b(x), \quad y=y(x), \quad x=ih\in\omega_h,$$
$$y_x=(y(x+h)-y(x))/h, \quad y_{\bar{x}}=(y(x)-y(x-h))/h.$$

Para que el problema (4) sea resoluble, es suficiente que sea $a>0$, $b>0$, $d\geqslant 0$, y en este caso la solución puede ser determinada por el método de factorización (véase el cap. I, § 3).

3. Condiciones de aproximación. Calculemos el error de aproximación del esquema (4):

$$\psi = (\Lambda v + \varphi) - (Lv + f) = (\Lambda v - Lv) + (\varphi - f) =$$
$$= \left[\frac{1}{h}(bv_x - av_{\bar{x}}) - (kv')' \right] - (d - q)v + (\varphi - f),$$

donde $v(x)$ es una función arbitraria suficientemente suave; k, q, f cuentan con un número de derivadas necesarias en el transcurso de la exposición. Hegamos uso de la fórmula de Taylor:

$$v(x \pm h) = v(x) \pm hv'(x) + \frac{h^2}{2}v''(x) \pm \frac{h^3}{6}v'''(x) + O(h^4)$$

y hallemos

$$v_x = v' + \frac{h}{2}v'' + \frac{h^2}{6}v''' + O(h^3),$$
$$v_{\bar{x}} = v' - \frac{h}{2}v'' + \frac{h^2}{6}v''' + O(h^3).$$

Sustituyamos estas expresiones para v_x y $v_{\bar{x}}$ en la fórmula para ψ:

$$\psi = \left(\frac{1}{h}(b - a) - k' \right) v' + \left(\frac{b + a}{2} - k \right) v'' +$$
$$+ \frac{h(b - a)}{6} v''' - (d - q)v + (\varphi - f) + O(h^2).$$

De aquí se ve que el esquema tiene el segundo orden de aproximación si quedan cumplidas las condiciones

$$\frac{b - a}{2} = k'(x) + O(h^2), \quad \frac{b + a}{2} = k(x) + O(h^2),$$
$$d = q(x) + O(h^2), \quad \varphi = f(x) + O(h^2). \quad (5)$$

En este caso $\psi = O(h^2)$.

El esquema (4) con los coeficientes

$$b_i = k_{i+1/2}, \quad a_i = k_{i-1/2}, \quad d_i = q_i, \quad \varphi_i = f_i,$$
$$b_1 = \frac{k_i + 2k_{i+1/2} + k_{i+1}}{4}, \; a_i = k_{i-1/2}, \quad d_i - q_i \quad \varphi_i = f_i,$$

satisface las condiciones (5) del segundo orden de aproximación, mientras que el esquema con los coeficientes

$$b_i = k_{i+1}, \quad a_i = \frac{k_i + k_{i+1}}{2}$$

no satisface ni siquiera la condición del primer orden de aproximación, puesto que

$$\frac{1}{h}(b_i - a_i) - k_i' = O(1).$$

§ 3. Esquemas de diferencias conservativos

1. Esquemas conservativos homogéneos. En el § 4, cap. I fue establecido que la condición necesaria y suficiente para que un operador de diferencias Λy sea autoconjugado (la matriz sea simétrica) consiste en que $b_i = a_{i+1}$. En este caso el problema (2) del § 2 adquiere la forma

$$\Lambda y = \frac{1}{h}\left[a_{i+1}\frac{y_{i+1}-y_i}{h} - a_i\frac{y_i - y_{i-1}}{h}\right] - d_i y_i = -\varphi_i,$$
$$i = 1, 2, \ldots, N-1, \quad y_0 = \mu_1, \quad y_N = \mu_2. \tag{1}$$

La ecuación

$$a_{i+1}\frac{y_{i+1}-y_i}{h} - a_i\frac{y_i - y_{i-1}}{h} - hd_i y_i = -h\varphi_i \tag{2}$$

es un análogo reticular de la ecuación de balance del calor sobre el intervalo $(x_{-1,2}, x_{i+1,2})$:

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx = -\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx, \quad w = ku',$$

(que se obtiene integrando la ecuación (1) del § 2 a lo largo del segmento $x_{i-1,2} \leqslant x \leqslant x_{i+1,2}$ y lleva el nombre de esquema *conservativo*, es decir, esquema para el cual se cumplen los análogos de diferencias de las leyes físicas de conservación.

El requisito $b_i = a_{i+1}$ para un esquema homogéneo significa que $B[k(x+sh)] = A[k(x+(s+1)/h]$, o bien, $B[\bar{k}(s)] = A[\bar{k}(s+1)]$ para cualesquiera funciones continuas a trozos $\bar{k}(s)$ en el segmento $[-1, 1]$. Esto es posi-

ble sólo cuando la funcional $A[\bar{k}(s)]$ no depende de los valores de $\bar{k}(s)$ para $0 \leqslant s \leqslant 1$, y $B[\bar{k}(s)]$ no depende de los valores de $\bar{k}(s)$ para $-1 \leqslant s \leqslant 0$, de modo que $a(x) = A[k(x+sh)]$ para $-1 \leqslant s \leqslant 0$. El coeficiente $a(x)$ del esquema conservativo depende sólo de los valores de $k(x)$ en el segmento $[x-h, x]$. Las condiciones del segundo orden de aproximación (5) del § 2 toman, para el esquema conservativo (2), la forma siguiente

$$\frac{a(x+h)-a(x)}{h} = k'(x) + O(h^2).$$
$$\frac{a(x+h)+a(x)}{2} = k(x) + O(h^2), \tag{3}$$

$$d(x) = q(x) + O(h^2), \quad \varphi(x) = f(x) + O(h^2). \tag{4}$$

De aquí, en particular, proviene que

$$a(x) = k(x) - {}^1/_2 hk'(x) + O(h^2) = k(x - {}^1/_2 h) + O(h^2).$$

Escribamos el esquema conservativo (2) utilizando las designaciones sin índices:

$$(ay_{\bar{x}})_x - d(x)\, y = -\varphi(x),$$
$$x = ih \in \omega_h, \quad y(0) = \mu_1, \quad y(1) = \mu_2. \tag{5}$$

Exigiremos que se cumplan también las condiciones

$$a \geqslant c_1 > 0, \qquad d \geqslant 0. \tag{6}$$

En la práctica se deben emplear las fórmulas sencillas para a, d y φ, por ejemplo, $a_i = k_{i-1/2}$, $d_i = q_i$, $\varphi_i = f_i$.

Si la discontinuidad de la función $k(x)$ se halla dentro del nodo $x = x_i$ de la red, calculemos los coeficientes del esquema homogéneo:

$$a_i = k_{i-1/2} \quad \text{o bien} \quad a_i = {}^1/_2 (k(x_{i-1}+0) + k(x_i - 0)),$$
$$d_i = {}^1/_2 (q(x_i - 0) + q(x_i + 0)), \quad \varphi_i = {}^1/_2 (f(x_i - 0) + f(x_i + 0)).$$

En este caso las condiciones (3) se cumplen en todo punto, mientras que las condiciones (4) se sustituyen por las condiciones

$$d_i - {}^1/_2 (q_{i-0} + q_{i+0}) = O(h^2), \qquad \varphi_i - {}^1/_2 (f_{i-0} + f_{i+0}) = O(h^2).$$

Demos a conocer los ejemplos de un esquema cuyos coeficientes se calculan por integración en los intervalos de la red:

$$a_i=\left(\frac{1}{h}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\right)^{-1}=\left(\int\limits_{-1}^{0}\frac{ds}{k(x_i+sh)}\right)^{-1};$$

$$\varphi_i=\frac{1}{h}\int\limits_{x_{i-1/2}}^{x_{i+1/2}}f(x)\,dx=\int\limits_{-1/2}^{1/2}f(x_i+sh)\,ds,$$

$$d_i=\frac{1}{h}\int\limits_{x_{i-1/2}}^{x_{i+1/2}}q(x)\,dx=\int\limits_{-1/2}^{1/2}q(x_i+sh)\,ds.$$

Es evidente, pues, que las condiciones (3), (4) se cumplen.

2. Error de aproximación. Veamos un esquema conservativo de segundo orden de la aproximación. Sea $u=u(x)$ la solución exacta del problema

$$Lu=(ku')'-q(x)\,u=-f(x),\quad 0<x<1,$$
$$u(0)=\mu_1,\quad u(1)=\mu_2, \tag{6}$$

y sea $y_i=y(x_i)$ la solución del problema de contorno en diferencias (5). Analicemos el error del esquema, es decir, una función reticular

$$z(x)=y(x)-u(x),\qquad x\in\bar{\omega}_h.$$

Al sustituir $y(x)=z(x)=u(x)$ en la ecuación (5) y al suponer que $u(x)$ es la función dada, obtendremos para el error $z(x)$ un problema

$$\Lambda z=(az_{\bar{x}})_x-dz=-\psi(x),\quad x\in\omega_h,$$
$$z(0)=0,\quad z(1)=0,\quad a\geqslant c_1>0,\quad d\geqslant 0, \tag{6'}$$

donde $\psi(x)=\Lambda u+\varphi(x)=(au_{\bar{x}})_x-du+\varphi$ es el residuo del esquema (5) en la solución $u=u(x)$ del problema diferencial de partida.

Teniendo presente que $Lu + f = 0$, escribamos

$$\psi = (\Lambda u + \varphi) - (Lu + f) = (\Lambda u - Lu) + (\varphi - f) =$$
$$= [(au_{\bar{x}})_x - (ku')'] - (d - q)\, u + (\varphi - f).$$

Por hipótesis, el esquema (5) satisface las condiciones del segundo orden de la aproximación. Esto significa que $\psi = O(h^2)$, si $k \in C^{(3)}$, $q, f \in C^{(2)}$, $u \in C^{(4)}$, y, por lo tanto,

$$\| \psi \|_C = O(h^2).$$

Con estas suposiciones el esquema tiene el segundo orden de exactitud.

No obstante, el mismo orden de exactitud tiene lugar también para las exigencias más débiles respecto a la suavidad:

$$k(x), \quad q(x), \quad f(x) \in C^{(2)}, \quad u \in C^{(3)}. \tag{7}$$

LEMA. *Si se cumplen las condiciones* (7), *queda lícita la fórmula*

$$\frac{(ku')_{i+1/2} - (ku')_{i-1/2}}{h} = (ku')'_i + O(h^2), \tag{8}$$

donde $u = u(x)$ *es la solución de la ecuación* (6).

DEMOSTRACIÓN. Hagamos uso de la fórmula de Taylor:

$$v_{i\pm 1/2} = v_i \pm \frac{1}{h} hv'_i + \frac{h^2}{8} v''_i + \frac{h^3}{48} v'''_i (x_i \pm \theta h),$$
$$0 \leqslant \theta \leqslant 1, \quad \frac{1}{h}(v_{i+1/2} - v_{i-1/2}) = v'_i + O(h_2).$$

Sustituyendo aquí $v = ku'$ y teniendo en cuenta que $(ku')'' = (qu - f)'$, $(ku')''' = (qu - f)''$, obtenemos la fórmula (8).

En virtud del lema, el error de aproximación ψ puede ser representado en la forma

$$\psi_i = \eta_{x,i} + \psi^*_i, \quad \eta_i = (au_{\bar{x}})_i - (ku')_{i-1/2}, \quad \psi^*_i = O(h^2)$$

bajo las condiciones (7).

Ahora, teniendo presente que

$$a_i = k_{i-1/2} + O(h^2) \quad \text{para} \quad k(x) \in C^{(2)},$$

$$u_{\bar{x},i} = \frac{u_i - u_{i-1}}{h} = (u')_{i-1/2} + O(h^2) \quad \text{para} \quad u \in C^{(3)},$$

obtenemos $\eta_i = O(h^2)$. Efectivamente, $u_i = u_{i-1/2} + {}^1/_2 h u'_{i-1/2} + {}^1/_8 h^2 u''_{i-1/2} + O(h^3)$.

$$u_{i-1} = u_{i-1/2} - \frac{1}{2} h u'_{i-1/2} + \frac{1}{8} h^2 u''_{i-1/2} + O(h^3),$$

$$u_{\bar{x},i} = u'_{i-1/2} + O(h^2),$$

$$a_i u_{\bar{x},i} = (k_{i-1/2} + O(h^2))(u'_{i-1/2} + O(h^2)) = (ku')_{i-1/2} + O(h^2),$$

$$\eta_i = O(h^2).$$

Más abajo se obtendrá la estimación apriorística $\| z \|_C$ directamente en términos de η y ψ^*.

3. Estimaciones apriorísticas. Pasemos a la estimación del error z en términos de ψ. Recordemos, ante todo, la estimación obtenida en el § 5 del cap. I con ayuda del método de factorización:

$$\| z \|_C \leqslant \frac{1}{c_1} \sum_{i=1}^{N-1} h \sum_{k=1}^{i} h |\psi_k|,$$

de donde se infiere

$$\| z \|_C \leqslant \frac{1}{2c_1} \| \psi \|_C.$$

Mostremos que para la solución del problema

$$(az_{\bar{x}})_x - dz = -\mu_x, \quad x \in \omega_h, \quad z(0) = z(1) = 0,$$
$$a \geqslant c_1 > 0, \qquad d \geqslant 0$$

es válida la estimación

$$\| z \|_C \leqslant \frac{2}{c_1} (1, |\mu|], \tag{9}$$

donde se designa $(y, v] = \sum_{i=1}^{N} y_i v_i h$.

Representemos z en forma de una suma $z = w + v$, donde w y v son las soluciones de los problemas

$$(aw_{\bar{x}})_x = -\mu_x, \quad x \in \omega_h, \quad w(0) = w(1) = 0;$$

$$\Lambda v = (av_{\bar{x}})_x - dv = -dw, \quad x \in \omega_h, \quad v(0) = v(1) = 0$$

La función w se hallará en la forma explícita, para estimar v se hará uso del principio del máximo. De la ecuación

$$(aw_{\bar{x}} + \mu)_x = 0, \quad (aw_{\bar{x}})_{i+1} = \mu_{i+1} = (aw_{\bar{x}})_i + \mu_i$$

se deduce que $aw_{\bar{x}} + \mu = \text{const} = c_0$. Realicemos las transformaciones evidentes:

$$w_i = w_{i-1} + \frac{(c_0 - \mu_i)\,h}{a_i} = c_0 \sum_{k=1}^{i} \frac{h}{a_k} - \sum_{k=1}^{i} \frac{\mu_k}{a_k} h + w_0,$$

$$0 = w_N = c_0 \sum_{k=1}^{N} \frac{h}{a_k} - \sum_{k=1}^{N} \frac{\mu_k}{a_k} h;$$

$$c_0 = \sum_{k=1}^{N} \frac{\mu_k}{a_k} h \Big/ \sum_{k=1}^{N} \frac{h}{a_k}.$$

Al introducir la designación

$$\alpha_i = \sum_{k=1}^{i} \frac{h}{a_k} \Big/ \sum_{k=1}^{N} \frac{h}{a_k}, \qquad 0 < \alpha_i \leqslant 1,$$

encontramos

$$w_i = \alpha_i \sum_{k=1}^{N} \frac{h\mu_k}{a_k} - \sum_{k=1}^{i} \frac{h\mu_k}{a_k}.$$

De aquí proviene

$$|w_i| = \left| -(1-\alpha_i) \sum_{k=1}^{i} \frac{h\mu_k}{a_k} + \alpha_i \sum_{k=i+1}^{N} \frac{h\mu_k}{a_k} \right| \leqslant$$

$$\leqslant (1-\alpha_i) \sum_{k=1}^{i} \frac{h\,|\mu_k|}{a_k} + \alpha_i \sum_{k=i+1}^{N} \frac{h\,|\mu_k|}{a_k} \leqslant \sum_{k=1}^{N} \frac{h\,|\mu_k|}{a_k}.$$

Ahora nos queda tomar en consideración que $a_k \geqslant c_1 > 0$ y obtenemos

$$\| w \|_C \leqslant \frac{1}{c_1} \sum_{k=1}^{N} h|\mu_k| = \frac{1}{c_1} (1, \ |\mu|). \tag{10}$$

Con el fin de estimar v hagamos uso del teorema 4 del § 5 del cap. I:

$$\| v \|_C \leqslant \| w \|_C. \tag{11}$$

Al reunir las desigualdades (10 y (11), tenemos

$$\| z \|_C = \| w + v \|_C \leqslant 2 \| w \|_C \leqslant \frac{2}{c_1} (1, |\mu|],$$

es decir, queda demostrada la estimación (9).

Volvamos ahora al problema (6′), donde $\psi = \eta_x + \psi^*$. Representemos en la forma

$$\psi = \mu_x, \quad \text{donde } \mu_i = \eta_i + \sum_{k=1}^{i-1} h\psi_k^*, \tag{12}$$

y hagamos uso de la estimación (9). Entonces, para la solución del problema (6′) obtendremos las siguientes estimaciones apriorísticas:

$$\begin{gathered} \| z \|_C \leqslant \frac{2}{c_1} \left\{ (1, |\eta|] + \sum_{k=1}^{N} h \left| \sum_{k=1}^{i-1} h\psi_k^* \right| \right\}, \\ \| z \|_C \leqslant \frac{2}{c_1} \{(1, |\eta|] + (1, |\psi^*|]\}. \end{gathered} \tag{13}$$

Queda probar que tiene lugar la fórmula (12). En efecto, al designar $\rho_i = \sum_{k=1}^{i-1} h\psi_k^*$, vemos que $\rho_{i+1} - \rho_i = h\psi_i^*$, es decir, $\psi_i^* = \rho_{x,i}$ y $\psi = \eta_x + \rho_x = \mu_x$, donde $\mu_i = \eta_i + \rho_i$.

4. Convergencia y exactitud del esquema de diferencias. Pasemos a estimar la exactitud de un esquema de diferencias. Suponiendo que

$$k(x), \quad q(x), \quad f(x) \in C^{(2)}, \quad u(x) \in C^{(3)},$$

obtenemos $\eta(x) = O(h^2)$, $\psi^* = O(h^2)$. Ahora resta por utilizar la estimación apriorística (13), la que podría ser sustituida por una estimación más aproximada

$$\| z \|_C \leqslant \frac{2}{c_1} (\| \eta \|_C + \| \psi^* \|_C).$$

De aquí se desprende que el esquema (5) converge uniformemente con el segundo orden, es decir, $\| z \|_C = \| y - u \|_C \leqslant Mh^2$, si se cumplen las condiciones (7).

Resulta más difícil demostrar la convergencia del esquema en la clase de coeficientes discontinuos $k(x)$, $q(x)$, $f(x)$. Para simplificar, analicemos un caso en que $k(x)$ tiene la discontinuidad de primera especie en un punto, mientras que $q(x)$ y $f(x)$ son continuas y pertenecen ambas a la clase $C^{(2)}$.

Denotemos con $Q^{(k)}[a, b]$ un conjunto de funciones continuas a trozos que están definidas en el segmento $[a, b]$ y tienen en $[a, b]$ k derivadas continuas a trozos.

Así pues, sea $k(x) \in Q^{(2)}$, $q(x)$, $f(x) \in C^{(2)}$ y $k(x)$ tiene discontinuidad de primera especie en el punto ξ del segmento $[x_n, x_{n+1}]$, de modo que $\xi = x_n + \theta h$, $0 \leqslant \theta \leqslant 1$. Para $x = \xi$ se cumplen las condiciones de conjugación

$$u_- = u_+, \quad (ku')_- = (ku')_+ = w_0,$$

donde

$$v_+ = v(\xi + 0), \quad v_- = v(\xi - 0).$$

Entonces $\eta_i = O(h^2)$ para $i \neq n+1$, $\psi_i^* = O(h^2)$ para todo $i = 1, 2, \ldots, N-1$, $\eta_{n+1} = a_{n+1}u_{x,n} - (ku')_{n+1/2}$. Sustituyendo aquí

$$u_{n+1} = u(\xi) + (1-\theta)hu'_+ + O(h^2),$$
$$u_n = u(\xi) - \theta hu'_- + O(h^2),$$

$$u_{x,n} = (u_{n+1} - u_n)/h = \theta u'_- + (1-\theta)u'_+ + O(h) =$$
$$= \theta\frac{(ku')_-}{k_-} + (1-\theta)\frac{(ku')_+}{k_+} + O(h) =$$
$$= w_0\left(\frac{\theta}{k_-} + \frac{1-\theta}{k_+}\right) + O(h),$$

$$(ku')_{n+1/2} = (ku')_- + O(h) = w_0 + O(h) \text{ para } \theta > 1/2,$$
$$(ku')_{n+1/2} = (ku')_+ + O(h) = w_0 + O(h) \text{ para } \theta < 1/2,$$

obtenemos

$$\eta_{n+1} = w_0\left[a_{n+1}\left(\frac{\theta}{k_-} + \frac{1-\theta}{k_+}\right) - 1\right] + O(h),$$

es decir, $\eta_{n+1} = O(1)$ para cualquier esquema y sólo para un esquema con coeficiente

$$\dot{a}_i = \left[\frac{1}{h}\int\limits_{x_{i-1}}^{x_i}\frac{dx}{k(x)}\right]^{-1}$$

tenemos $\eta_{n+1} = O(h)$. En efecto,

$$\frac{1}{\overset{\bullet}{a}_{n+1}} = \frac{1}{h}\int_{x_n}^{\xi}\frac{dx}{k(x)} + \frac{1}{h}\int_{\xi}^{x_{n+1}}\frac{dx}{k(x)} = \frac{\theta}{k_-} + \frac{1-\theta}{k_+} + O(h),$$

es decir, $\overset{\bullet}{a}_{n+1}\left(\frac{\theta}{k_-} + \frac{1-\theta}{k_+}\right) = 1 + O(h)$, y, por consiguiente, $\eta_{n+1} = O(h)$. En el segundo miembro de la desigualdad (13) figura la magnitud

$$(1,\ |\eta|] = \sum_{i=1,\ i\neq n+1}^{N} h|\eta_i| + h|\eta_{n+1}|.$$

Con esto queda demostrado el teorema siguiente.

TEOREMA. En *la clase de coeficientes discontinuos* $k(x) \in \in Q^2$, $q(x)$, $f(x) \in C^{(2)}$ *cualquter esquema de diferencias homogéneo* (5) *de segundo orden de la aproximación tiene el primer orden de exactitud, mientras que el esquema con coeficiente* $a_i = \overset{\bullet}{a}_i$ *tiene el segundo orden de exactitud.*

§ 4. Esquemas homogéneos sobre las redes no uniformes

1. Esquema conservativo en una red no uniforme. Elijamos en el segmento $0 \leqslant x \leqslant 1$ una red arbitraria no uniforme

$$\hat{\omega}_h = \{x_i,\ i = 0,\ 1,\ \ldots,\ N,\ x_0 = 0,\ x_N = 1\}.$$

Para obtener un esquema conservativo tripuntual en la red no uniforme, escribamos una ecuación de balance en el segmento $[x_{i-1/2},\ x_{i+1/2}]$:

$$w_{i+1/2} - w_{i-1/2} - \int_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx = -\int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx,\quad w = ku'.$$

Dicha ecuación se anota igual tanto para una red uniforme como para una no uniforme. Nos queda aproximar las inte-

grales y derivadas que intervienen en la ecuación de balance:

$$w_{i-1/2}=(ku')_{i-1/2}\sim a_i\frac{u_i-u_{i-1}}{h_i},\quad h_i=x_i-x_{i-1},$$

$$\int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx\sim\varphi_i\hbar_i,\quad \int\limits_{x_{i-1/2}}^{x_{i+1/2}} qu\,dx\sim d_iu_i\hbar_i,$$

$$\hbar_i=\frac{1}{2}(h_i+h_{i+1}),$$

donde d_i y φ_i son unas funciones reticulares. Como resultado, se obtiene un esquema de diferencias

$$\frac{1}{\hbar_i}\left[a_{i+1}\frac{y_{i+1}-y_i}{h_{i+1}}-a_i\frac{y_i-y_{i-1}}{h_i}\right]-d_iy_i=-\varphi_i,$$
$$i=1,\ 2,\ \ldots,\ N-1,\quad y_0=\mu_1,\quad y_N=\mu_2. \quad (1)$$

Para determinar d_i y φ_i usaremos las fórmulas más sencillas $\varphi_i=f_i$, $d_i=q_i$, $i=1,\ 2,\ \ldots,\ N-1$. El coeficiente a_i, se determina por los valores $k(x)$ en el intervalo $(x_{i-1},\ x_i)$, a consecuencia de lo cual puede tomarse igual al que figura sobre la red uniforme, de modo que $a_i=k_{i-1/2}+O(h_i^2)$ para $k(x)\in C^{(2)}$.

2. Error de aproximación. Introduzcamos las designaciones

$$y_{\bar{x},i}=\frac{y_i-y_{i-1}}{h_i},\quad y_{x,i}=\frac{y_{i+1}-y_i}{h_{i+1}},\quad y_{\hat{x},i}=\frac{y_{i+1}-y_i}{\hbar_i}$$

y escribamos el esquema de diferencias en la forma

$$(ay_{\bar{x}})_{\hat{x}}-dy=-\varphi,\ x=x_i\in\hat{\omega}_h,\ y_0=\mu_1,\ y_N=\mu_2. \quad (1)$$

Al suponer $z=y-u$, obtendremos para z la ecuación

$$(az_{\bar{x}})_{\hat{x}}-dz=-\psi,\quad x\in\hat{\omega}_h,\quad z_0=z_N=0, \quad (2)$$

donde

$$\psi=\Lambda u+\varphi=(au_{\bar{x}})_{\hat{x}}-du+\varphi \quad (3)$$

es el residuo para el esquema (1) en la solución $u=u(x)$.

LEMA 1. *Si* $qu \in C^{(2)}$, $f \in C^{(2)}$, *entonces para el error de aproximación* ψ *es válida la fórmula*

$$\psi = \eta_{\hat{x}} + \psi^*, \tag{4}$$

donde $\eta_i = (au_{\bar{x}})_i - (ku')_{i-1/2} - h_i^2 (qu - f)'_i/8$, $\psi_i^* = O(h^2)$ *para* $\varphi_i = f_i$, $d_i = q_i$.

Hagamos uso de la indentidad del p. 1, escribiéndola en la la forma

$$0 = w_{\hat{x},i} - \frac{1}{\hbar_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} (qu - f)\,dx, \quad w_i = (ku')_{i-1/2}.$$

Sustrayemos esta identidad de la igualdad (3):

$$\psi = [(au_{\bar{x}})_i - (ku')_{i-1/2}] - (du)_i + \varphi_i + \frac{1}{\hbar_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} (du - f)\,dx. \tag{5}$$

La integral que figura en el segundo miembro se representará en forma de una suma de dos integrales: de $x_{i-1/2}$ a x_i y de x_i a $x_{i+1/2}$; al desarrollar después la función subintegral $\tilde{f} = qu - f$ en el entorno del nodo $x = x_i$, hallaremos

$$\frac{1}{\hbar_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x)\,dx = \frac{1}{\hbar_i} \Bigg\{ \int\limits_{x_{i-1/2}}^{x_i} [\tilde{f}_i + (x - x_i)\tilde{f}'_i]\,dx + O(h_i^3) +$$

$$+ \int\limits_{x_i}^{x_{i+1/2}} [\tilde{f}_i + (x - x_i)\tilde{f}'_i]\,dx + O(h_{i+1}^3) \Bigg\} =$$

$$= \tilde{f}_i + \frac{1}{8\hbar_i}(h_{i+1}^2 - h_i^2)\tilde{f}'_i + O(\hbar_i^2),$$

puesto que $h_i^3 + h_{i+1}^3 < (2\hbar_i)^3$. La sustitución $h_{i+1}^2 \tilde{f}'_i = h_{i+1}^2 \tilde{f}'_{i+1} + O(h_{i+1}^3)$ nos da

$$\frac{1}{\hbar_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} \tilde{f}(x)\,dx = \tilde{f}_i + (h^2 \tilde{f}')_{\hat{x},i} + O(\hbar_i^2).$$

Sustituyendo esta expresión con $\tilde{f} = qu - f$ en (5), llegamos a la fórmula (4).

Para estimar η_i según el orden veamos la diferencia $(au_{\bar{x}})_i - (ku')_{i-1/2}$ a condición de que $k \in C^{(2)}$, $u \in C^{(3)}$. Empleando la suposición $a_i = k_{i-1/2} + O(h_i^2)$ y las fórmulas $u_i = u_{i-1/2} + h_i u'_{i-1/2}/2 + h_i^2 u''_{i-1/2}/8 + O(h_i^3)$, $u_{i-1} = u_{i-1/2} - h_i u'_{i-1/2}/2 + h_i^2 u''_{i-1/2}/8 + O(h_i^3)$, $u_{\bar{x},i} = (u_i - u_{i-1})/h_i = u'_{i-1/2} + O(h_i^2)$, obtenemos

$$(au_{\bar{x}})_i - (ku')_{i-1/2} =$$
$$= (k_{i-1/2} + O(h_i^2))(u'_{i-1/2} + O(h_i^2)) - (ku')_{i-1/2} = O(h_i^2).$$

De este modo, es válida la estimación

$$\eta_i = O(h_i^2) \text{ para } (k(x),\ q(x),\ f(x) \in C^{(2)},\ u(x) \in C^{(3)}.$$

OBSERVACIÓN. Se suponía que d_i y φ_i se determinan según las fórmulas más sencillas: $d_i = q_i$, $\varphi_i = f_i$. Si, en cambio, se emplean las fórmulas más complejas, por ejemplo

$$\varphi_i = \frac{h_i f_{i-1/2} + h_{i+1} f_{i+1/2}}{2\hbar_i}, \quad \varphi_i = \frac{1}{\hbar_i} \int\limits_{x_{i-1/2}}^{x_{i+1/2}} f(x)\, dx,$$

entonces la función reticular $\psi_i^* = O(\hbar_i^2) - (d_i - q_i) u_i + (\varphi_i - f_i)$ puede ser representada en la forma $\psi_i^* = \rho_{\hat{x},i} + \psi_i^{**}$, donde $\psi_i^{**} = O(\hbar_i^2)$, $\rho_i = O(\hbar_i^2)$ y η_i, en la fórmula (4) se sustituye por la suma $\eta_i + \rho_i$:

$$\psi = (\rho_i + \eta_i)_{\hat{x}} + \psi^{**}, \tag{4'}$$
$$\rho_i = O(h_i^2) \quad \eta_i = O(h_i^2), \quad \psi_i^{**} = O(\hbar_i^2)$$
$$\text{para} \quad k,\ q,\ f \in C^{(2)}, \quad u \in C^{(3)}.$$

3. Estimación de la velocidad de convergencia. Para el problema (2)-(4) es válida la siguiente estimación apriorística

$$\| z \|_C \leqslant \frac{1}{c_1} \{ (1,\ |\eta|] + (1,\ |\psi^*|] \}, \tag{6}$$

donde $(y,\ v)] = \sum_{i=1}^{N} y_i v_i h_i$. Si se cumplen las condiciones (7) del § 3, entonces $\eta_i = O(h_i^2)$, $\psi_i^* = O(\hbar_i^2)$.

Al sustituir η_i y ψ_i^* en (6), nos convencemos de que es verídico el siguiente teorema.

TEOREMA. *En la clase de coeficientes suaves $k, q, f \in C^{(2)}$ todo esquema de la forma (1) mantiene el segundo orden de exactitud en una sucesión arbitraria de las redes no uniformes.*

Al tomar en consideración la observación del p. 2, podemos representar ψ_i^* en la forma $\psi_i^* = \rho_{\hat{x},i} + \psi_i^{**}$, donde $\rho_i = O(h_i^2)$, $\psi_i^{**} = O(\hbar_i^2)$. Entonces, en lugar de (6) queda válida la estimación

$$\| z \|_C \leqslant \frac{2}{c_1} \{(1, |\eta + \rho|] + (1, |\psi^{**}|]\};$$

el teorema sobre el segundo orden de exactitud sobre una red no uniforme queda en vigor.

Si el coeficiente $k(x)$ tiene discontinuidades de primera especie en un número finito de puntos, siempre podemos elegir tal red no uniforme $\hat{\omega}_h(k)$ que los puntos de discontinuidad sean los nodos de dicha red. En tal caso cualquier esquema tendrá el segundo orden de exactitud.

Así pues, cualquier esquema homogéneo de segundo orden de aproximación ($\psi = O(h^2)$) sobre una red no uniforme y en la clase de coeficientes suaves tiene el segundo orden de exactitud con la elección especial de las redes no uniformes $\hat{\omega}_h(k)$ en la clase de coeficientes discontinuos.

4. Esquema exacto. Para el problema (1) del § 2 podemos construir un esquema tripuntual exacto cuya solución en los nodos de una red arbitraria coincide con la solución exacta $u = u(x)$ del problema de contorno para una ecuación diferencial. Ilustremos la posibilidad de construir el esquema exacto con un caso particular del problema para $q(x) \equiv 0$:

$$(ku')' = -f(x), \quad 0 < x < 1, \quad u(0) = 0, \quad u(1) = 0. \quad (7)$$

Al haber integrado la ecuación desde x_i hasta x, obtendremos una ecuación

$$(ku') - (ku')_i + \int_{x_i}^{x} f(\xi)\, d\xi = 0.$$

Dividámosla por $k(x)$ y integremos respecto de x desde x_i hasta x_{i+1}:

$$u_{i+1} - u_i - (ku')_i \int_{x_i}^{x_{i+1}} \frac{dx}{k(x)} + \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi)\, d\xi = 0, \quad (8)$$

y, luego, de x_{i-1} a x_i:

$$u_i - u_{i-1} - (ku')_i \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)} + \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(\xi)\, d\xi = 0. \qquad (9)$$

Introduzcamos una designación

$$a_i^0 = \left[\frac{1}{h_i} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\right]^{-1}.$$

Multipliquemos (8) por a_{i+1}^0/h_{i+1}, (9) por a_i^0/h_i, y sustrayamos del primer resultado el segundo. Obtendremos una ecuación

$$\frac{1}{\hbar_i}\left[a_{i+1}^0 \frac{u_{i+1}-u_i}{h_{i+1}} - a_i^0 \frac{u_i - u_{i-1}}{h_i}\right] + \varphi_i = 0,$$

o bien

$$(a^0 u_{\bar{x}})_{\hat{x},\, i} + \varphi_i = 0, \qquad (10)$$

donde

$$\varphi_i = \frac{a_i^0}{h_i \hbar_i} \int_{x_{i-1}}^{x_i} \frac{dx'}{k(x')} \int_{x'}^{x_i} f(t)\, dt + \frac{a_{i+1}^0}{h_{i+1}\hbar_i} \int_{x_i}^{x_{i+1}} \frac{dx'}{k(x')} \int_{x_i}^{x'} f(t)\, dt.$$

Si ponemos $x' = x_i + sh_i$ para $x_{i-1} \leqslant x' \leqslant x_i$, y $x' = x_i + sh_{i+1}$ para $x_i \leqslant x' \leqslant x_{i+1}$, entonces dicha fórmula puede reescribirse de la manera siguiente:

$$\varphi_i = \frac{h_i a_i^0}{\hbar_i} \int_{-1}^{0} \frac{ds}{k(x_i + sh_i)} \int_{s}^{0} f(x_i + \lambda h_i)\, d\lambda +$$

$$+ \frac{h_{i+1} a_{i+1}^0}{\hbar_i} \int_{0}^{1} \frac{ds}{k(x_i + sh_{i+1})} \int_{0}^{s} f(x_i + \lambda h_{i+1})\, d\lambda.$$

De este modo, el esquema (10) es exacta sobre una red arbitraria no uniforme y para cualesquiera funciones continuas a trozos $k(x)$ y $f(x)$. Por supuesto, el empleo práctico de este esquema está obstaculizado por el hecho de que los coeficientes de dicho esquema se expresan a través de las in-

tegrales de $k(x)$ y $f(x)$, razón por la cual su cálculo requiere la aplicación de las fórmulas de integración numérica.

5. **Aumento del orden de exactitud.** De lo dicho se hace claro que para aumentar la exactitud de la solución aproximada se debe o bien disminuir el paso de la red h o bien aumentar el orden de exactitud del esquema. No obstante, es conveniente construir esquemas con orden de exactitud aumentado sólo para las ecuaciones de coeficientes constantes, puesto que la anotación de tales esquemas para las ecuaciones de coeficientes variables está relacionada con grandes dificultades técnicas y conduce, a menudo, a los algoritmos engarrosos. Ya hemos aducido un ejemplo del esquema $O(h^4)$ para la ecuación $u'' = -f(x)$.

Examinemos ahora una ecuación

$$u'' - qu = -f(x), \qquad q = \text{const} > 0.$$

Escribamos un esquema de diferencias sobre la red uniforme:

$$\Lambda y = y_{\bar{x}x} - dy = -\varphi(x)$$

y elijamos d y φ de un modo tal que tenga la aproximación $O(h^4)$. El error de la aproximación es

$$\psi = \Lambda u + \varphi = (\Lambda u - u'') - (d - q)u + \varphi - f =$$
$$= \frac{h^2}{12}u^{\text{IV}} - (d - q)u + \varphi - f + O(h^4).$$

Al sustituir aquí $u^{\text{IV}} = qu'' - f'' = q(qu - f) - f'' = q^2u - qf - f''$, obtendremos

$$\varphi = -\left(d - q - \frac{h^2}{12}q^2\right)u + \varphi - \left(f + \frac{h^2}{12}qf + \frac{h^2}{12}f''\right) + O(h^4);$$

por consiguiente, $\psi = O(h^4)$, si se pone $d = q + \frac{h^2}{12}q^2$, $\varphi = f + \frac{h^2}{12}(qf + f'')$. El orden de exactitud queda intacto, si en la fórmula para φ sustituimos la derivada f'' por su aproximación de diferencias $f_{\bar{x}x}$, puesto que $h^2f'' = h^2f_{\bar{x}x} + O(h^4)$.

El aumento de exactitud del esquema disminuyendo h viene limitada también por el requisito de la economía del tiempo indispensable para la obtención de la solución con

una exactitud prefijada. Por ello, en la práctica se utiliza con frecuencia el cálculo según un mismo esquema sobre la sucesión de redes, el cual permite elevar la exactitud sin aumentar considerablemente el tiempo de cálculo (método de Runge), bajo el supuesto de que sea la solución lo suficientemente suave.

Supongamos que para resolver un problema de diferencias en cualquier red uniforme es válido un desarrollo asintótico

$$y_i^h = u_i + \alpha(x_i)\, h^{k_1} + O(h^{k_2}), \quad k_2 > k_1 > 0, \tag{11}$$

donde $\alpha(x_i)$ no depende de h. Se pide hallar una función reticular $\tilde{y}_i$, para la cual

$$\tilde{y}_i = u_i + O(h^2) \tag{12}$$

sobre cierto conjunto de nodos $\tilde{\omega}_h$.

Veamos dos redes ω_{h_1} y ω_{h_2} de pasos h_1 y h_2, respectivamente, que tienen nodos comunes; designemos con $\tilde{\omega}_h$ el conjunto de nodos comunes. Sean $y_i^{h_1}$ e $y_i^{h_2}$ las soluciones del problema de diferencias en las redes ω_{h_1} y ω_{h_2}, respectivamente. Formemos su combinación lineal $\tilde{y}_i = \sigma y_i^{h_1} + (1-\sigma) y_i^{h_2}$ y sustituyamos aquí el desarrollo (11):

$$\tilde{y}_i = u_i + \alpha(x_i)\,(\sigma h_1^{k_1} + (1-\sigma)\, h_2^{k_1} + O(h^{k_2}).$$

Igualando a cero el coeficiente de $\alpha(x_i)$, hallemos

$$\sigma = h_2^{k_1}/(h_2^{k_1} - h_1^{k_1}); \tag{13}$$

con la particularidad de que en los nodos $x_i \in \tilde{\omega}_h$ se cumple el requisito (12).

De este modo, con el fin de aumentar la exactitud de la solución reticular en cierto conjunto de nodos $\tilde{\omega}_h$, se debe resolver el problema dos veces sobre las redes ω_{h_1} y ω_{h_2}, que se intersecan en dicho conjunto, y formar su combinación lineal con coeficientes σ y $(1-\sigma)$, donde σ se determina de acuerdo con (13).

En particular, podemos tomar $h_2 = h_1/2$, $h_1 = h$; entonces $\tilde{\omega}_h = \omega_{h_1}$. Para el esquema de segundo orden de exactitud tenemos $k_1 = 2$, $k_2 = 4$, y $\sigma = -1/3$, $1 - \sigma = 4/3$.

La posibilidad de obtener el desarrollo

$$z_i = y_i - u_i = \alpha(x_i)h^2 + O(h^4)$$

proviene del desarrollo del residuo $\psi_i = \beta(x_i)h^2 + O(h^4)$, el cual constituye el segundo miembro del problema

$$\Lambda z = -\psi, \qquad z_0 = z_N = 0.$$

El empleo de las redes no uniformes concede grandes posibilidades de aumento empírico de la exactitud sin aumentar el número de nodos, siempre que se tiene una información preliminar sobre el comportamiento de la solución del problema de partida. Así, en la región de variación fuerte de los *coeficientes y del segundo miembro de la ecuación resulta* natural espesar la red. Cerca de la frontera de una discontinuidad de los coeficientes, la red se espesa corrientemente según la ley de una progresión geométrica. Para obtener la información preliminar, se pueden realizar los primeros cálculos en una red aproximada y a continuación, los cálculos definitivos en una red especial.

§ 5. Métodos de construcción de los esquemas de diferencias

De lo expuesto más arriba está claro que los esquemas de diferencias para una ecuación diferencial concreta han de reflejar correctamente, en el espacio de funciones reticulares, las propiedades principales del problema de partida (autoconjugación, definición de signo y otras). Para el problema de contorno analizado por nosotros anteriormente, el requisito principal resultó ser una propiedad de conservación que es equivalente a la propiedad de autoconjugación del operador de diferencias. El problema de importancia consiste en obtener los esquemas de diferencias con una calidad prefijada. Para construir tales esquemas se emplean actualmente toda una serie de métodos, de los cuales se trata en este párrafo.

1. Método integral de interpolación. Una ecuación diferencial expresa habitualmente cierta ley física de conservación. Dicha ley puede ser escrita en la forma integral para *un intervalo (célula) de una red (ecuación de balance)*. La ecuación diferencial se obtiene de la ecuación de balance,

cuando el paso de la red tiende a cero bajo el supuesto de que existen derivadas continuas que figuran en la ecuación. Las derivadas e integrales que intervienen en la ecuación de balance sobre la red se deben sustituir por las expresiones aproximadas en la red. De resultas se obtendrá un esquema homogéneo. Este método se denomina *integral de interpolación* o bien *método de balance*. Ilustrémoslo con un ejemplo de un problema

$$(ku')' - qu = -f(x), \quad 0 < x < 1, \quad (ku') - \sigma_1 u = -\mu_1 \text{ para } x = 0, \quad u(1) = \mu_2. \tag{1}$$

Escribamos la ecuación de balance del calor en el segmento $0 \leqslant x \leqslant 1$.:

$$w_{i+1/2} - w_{i-1/2} + \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\,dx = \int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,u(x)\,dx, \quad w = ku', \tag{2}$$

donde $(-w(x))$ es el flujo térmico, $q(x)\,u(x)$ es la potencia de las corrientes (de las fuentes, cuando $q < 0$) de calor, la cual es proporcional a la temperatura, y $f(x)$, la densidad de distribución de las fuentes exteriores (de las corrientes) de calor. En el primer miembro de esta ecuación figura la cantidad de calor que queda a cuenta de los flujos térmicos en el segmento $[x_{i-1/2}, x_{i+1/2}]$ y a cuenta de las fuentes exteriores; en el segundo miembro se indica la cantidad de calor que se disipa al ambiente exterior a cuenta del intercambio térmico en la superficie lateral.

Con el objeto de obtener de (2) una ecuación en diferencias tripuntual, sustituyamos $w_{i-1/2}$, $w_{i+1/2}$ y las integrales en la ecuación (2), por la combinación lineal de valores de las funciones subintegrales en los nodos de la red (x_{i-1}, x_i, x_{i+1}), por ejemplo,

$$\frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,u(x)\,dx \approx d_i u_i, \quad d_i = \frac{1}{h}\int_{x_{i-1/2}}^{x_{i+1/2}} q(x)\,dx.$$

Integremos la igualdad $u' = w/k$ respecto de x entre x_{i-1} y x_i:

$$u_i - u_{i-1} = \int_{x_{i-1}}^{x_i} \frac{w}{k(x)}\, dx \approx h w_{i-1/2} \frac{1}{a_i},$$

$$a_i = \left[\frac{1}{h} \int_{x_{i-1}}^{x_i} \frac{dx}{k(x)}\right]^{-1}.$$

Como resultado, obtenemos de (2) un esquema

$$\frac{1}{h}\left[a_{i+1} \frac{y_{i+1} - y_i}{h} - a_i \frac{y_i - y_{i-1}}{h}\right] - d_i y_i = -\varphi_i,$$

$$\varphi_i = \frac{1}{h} \int_{x_{i-1/2}}^{x_{i+1/2}} f(x)\, dx.$$

Deduciendo esta expresión se suponía, de hecho, que $u = \text{const}$ para $x_{i-1/2} \leqslant x \leqslant x_{i+1/2}$, $w = \text{const}$ para $x_{i-1} \leqslant x \leqslant x_i$.

En lugar de las expresiones para a_i, d_i, φ_i conviene tomar las fórmulas más sencillas, como lo hicimos en los párrafos anteriores. Escribamos una aproximación de diferencias para la condición de contorno de tercera especie cuando $x = 0$. Con este fin hagamos uso de la ecuación de balance para $0 \leqslant x \leqslant x_{1/2} = h/2$

$$w_{1/2} - w_0 - \int_0^{x_{1/2}} qu\, dx = -\int_0^{x_{1/2}} f(x)\, dx.$$

Sustituyendo aquí

$$w_{1/2} = a_1 u_{\bar{x},1}, \qquad w_0 = (ku')_0 = \sigma_1 u_0 - \mu_1,$$

$$\int_0^{x_{1/2}} qu\, dx \sim q_0 u_0 \frac{1}{2} h, \qquad \int_0^{x_{1/2}} f(x)\, dx \sim f_0 \frac{1}{2} h$$

y cambiando en todos los casos u por y, obtendremos la condición de contorno en diferencias

$$a_1 y_{\bar{x},1} - \sigma_1 y_0 + \mu_1 - h q_0 y_0/2 = -h f_0/2.$$

la cual puede escribirse en la forma

$$a_1 y_{\bar{x},1} = \bar{\sigma}_1 y_0 - \bar{\mu}_1, \text{ donde } \bar{\sigma}_1 = \sigma_1 + hq_0/2, \ \bar{\mu}_1 = \mu_1 + hf_0/2. \quad (3)$$

Estimemos en la solución $u = u(x)$ de la ecuación (1) el valor del residuo

$$\nu = a_1 u_{\bar{x},1} - \bar{\sigma} u_0 + \bar{\mu}_1.$$

Al sustituir aquí $a_1 = k_{1/2} + O(h^2) = k_0 + {}^1/_2\, hk_0' + O(h^2)$, $u_1 = u_0 + hu_0' + h^2 u_0''/2 + O(h^3)$, $u_{\bar{x},1} = (u_1 - u_0)/h = u_0' + hu_0''/2 + O(h^2)$, obtenemos

$$\nu = (ku')_0 + {}^1/_2 h\,(ku')_0' - \bar{\sigma}_1 u_0 + \bar{\mu}_1 + O(h^2) =$$
$$= [(ku')_0 - \sigma_1 u_0 + \mu_1) + {}^1/_2 h\,[(ku')' - qu + f]_0 + O(h^2) = O(h^2),$$

es decir, la condición de contorno en diferencias de tercera especie (3) aproxima la condición $ku' = \sigma_1 u - \mu_1$ para $x = 0$ con un error de segundo orden $\nu = O(h^2)$.

Para el empleo práctico la condición de contorno (3) ha de escribirse en la forma

$$y_0 = \varkappa_1 y_1 - \tilde{\mu}_1, \quad \varkappa_1 = \frac{a_1}{a_1 + h\hat{\sigma}_1}, \quad \tilde{\mu}_1 = \frac{h\bar{\mu}_1}{a_1 + h\bar{\sigma}_1}.$$

Con el fin de aumentar la exactitud del esquema al calcular las integrales se debe utilizar la interpolación de orden más elevado.

2. Método de aproximación de una funcional cuadrática. Un problema de contorno

$$Lu = (ku')' - qu = -f(x), \quad 0 < x < 1, \quad u(0) = 0, \quad u(1) = 0$$

es equivalente al problema de buscar el elemento minimizador de la funcional cuadrática

$$J[u] = \int_0^1 (k(u')^2 + qu^2\, dx - 2\int_0^1 fu\, dx.$$

Introduzcamos en el segmento $0 \leqslant x \leqslant 1$ una red $\bar{\omega}_h = \{x_i = ih,\ i = 0, 1, \ldots, N\}$ y aproximemos la funcional.

Con este objeto representémosla, al principio, como una suma de integrales en los intervalos de la red:

$$J[u]=\sum_{i=1}^{N} J_i[u], \quad J_i[u]=\int_{x_{i-1}}^{x_i} (k(u')^2+qu^2-2fu)\,dx,$$

después de lo cual aproximemos J_i, por ejemplo, así:

$$\int_{x_{i-1}}^{x_i} k(u')^2\,dx \approx a_i (u_{\bar{x}, i})^2 h,$$

$$\int_{x_{i-1}}^{x_i} (qu^2-2fu)\,dx \approx \frac{h}{2}[(qu^2-2fu)_i+(qu^2-2fu)_{i-1}],$$

donde a_i es un coeficiente, por ejemplo,

$$a_i=\frac{1}{h}\int_{x_{i-1}}^{x_i} k(x)\,dx.$$

Obtenemos, como resultado, una funcional

$$J_h[y]=\sum_{k=1}^{N} ha_k(y_{\bar{x},k})^2+\sum_{k=1}^{N-1} (q_k y_k^2-2f_k y_k)\,h,$$

donde $y_i = y(i)$ es una función reticular arbitraria que se reduce a cero cuando $i = 0, N$.

La ecuación

$$Ay=\varphi \text{ ó } \sum_{j=1}^{N} a_{ij}y_j=\varphi_i, \qquad A=A^*>0,$$

tiene una solución que minimiza la funcional

$$I_A[y]=(Ay, y)-2(\varphi, y)=\sum_{i,j=1}^{N} a_{ij}y_jy_i-2\sum_{i=1}^{N}\varphi_i y_i.$$

De esto podemos convencernos al igualar a cero la derivada

$$\frac{\partial I_A[y]}{\partial y_{i_0}}=2\sum_{j=1}^{N} a_{i_0}y_j-2\varphi_{i_0}=0, \qquad \frac{\partial^2 I_A}{\partial y_{i_0}^2}>0,$$

puesto que $a_{ii} > 0$ para cualesquiera $i = 1, 2, \ldots, N$, en virtud de que A es positivo ($A > 0$).

Al calcular las derivadas

$$\frac{\partial J_h}{\partial y_i} = 2a_i y_{\bar{x},i} - 2a_{i+1} y_{\bar{x},i+1} + (2q_i y_i - 2f_i) h,$$

$$\frac{\partial J_h}{\partial y_i^2} = \frac{2a_i}{h} + \frac{2a_{i+1}}{h} + 2q_i > 0,$$

nos cercioramos de que el elemento $y = y(x) \in H_h$, que minimiza la funcional cuadrática, es la solución del problema

$$(ay_{\bar{x}})_{x,i} - q_i y_i = -f_i, \quad i = 1, 2, \ldots, N-1, \quad y_0 = 0 = y_N = 0$$

3. Método de aproximación de una identidad integral (método de identidades sumadoras). Sea

$$(ku')' - qu + f(x) = 0, \quad 0 < x < 1, \quad u(0) = u(1) = 0. \quad (4)$$

Multiplicando la ecuación (4) por una función diferenciable arbitraria $v(x)$, que se anula para $x = 0$, $x = 1$, e integrando respecto de k entre 0 y 1, obtenemos una identidad

$$I(u, v) = \int_0^1 (ku'v' + quv - fv)\,dx = 0.$$

Cambiando, por analogía con el p. 2, la integral y las derivadas u', v', escribamos una identidad sumadora

$$I_h[y, v] = \sum_{i=1}^{N} a_i y_{\bar{x},i} v_{\bar{x},i} h + \sum_{i=1}^{N-1} (q_i y_i - f_i) v_i h = 0.$$

Luego, suponiendo, por ejemplo, que $v_i = \delta_{i,i_0}$, $0 < i_0 < N$, y teniendo presente que $v_{\bar{x},i} = 0$ para $i < i_0$ y $i > i_0 + 1$, $v_{\bar{x},i_0+1} = -1/h$, $v_{\bar{x},i_0} = 1/h$, obtendremos

$$h\left(\frac{1}{h} a_{i+1} y_{x,i} - \frac{1}{h} a_i y_{\bar{x},i}\right) + (q_i - f_i) h = 0 \quad \text{cuando} \quad i = i_0,$$

es decir, $(a_{\bar{x}} y)_x - qy = -f$.

4. Métodos de Ritz y de Bubnov—Galerkin (métodos variacionales de diferencias). El problema sobre el mínimo

de una funcional

$$I[u] = (Au, u) - 2(u, f),$$

donde A es un operador lineal autoconjugado y definido positivo en el espacio de Hilbert H con el producto escalar (x, y), es equivalente al problema sobre la resolución de una ecuación

$$Au = f.$$

Se introduce una sucesión de espacios de dimensión finita V_n con base $\{\varphi_i^{(n)}\}$, $i = 1, 2, \ldots, n$.

El *método de Ritz* consiste en que se busca un elemento $u_n \in V_n$ que minimiza la funcional $I(u)$ en V_n. La solución aproximada u_n se busca en forma de la suma

$$u_n = \sum_{j=1}^{n} y_j \varphi_j, \tag{5}$$

donde $y_1, \ldots, y_n$ son unos coeficientes desconocidos. Los cálculos nos dan

$$I[u_n] = \sum_{i, j=1}^{n} \alpha_{ij} y_i y_j - 2 \sum_{i=1}^{n} \beta_i y_i,$$

$$\alpha_{ij} = \alpha_{ji} = (A\varphi_i, \varphi_j), \quad \beta_i = (f, \varphi_i);$$

$I[u_n] = \Phi(y_1, y_2, \ldots, y_n)$ es una función de n coeficientes y_i. Igualando a cero las derivadas $\partial I[u_n]/\partial y_i$, obtendremos un sistema de n ecuaciones

$$\sum_{j=1}^{n} \alpha_{ij} y_j - \beta_i = 0, \quad i = 1, 2, \ldots, n,$$

para determinar $y_1, y_2, \ldots, y_n$.

Ilustremos el método de Ritz con un ejemplo del problema (4). A título de la función $\varphi_i(x)$ tomemos

$$\varphi_i(x) = \eta\left(\frac{x - x^i}{h}\right) = \eta_i(x), \quad \eta(s) \begin{cases} 0, & s < -1, \ s > 1, \\ 1 + s, & -1 < s < 0, \\ 1 - s, & 0 \leqslant s \leqslant 1. \end{cases}$$

Al sustituir en la fórmula para $\alpha_{ij} A\varphi_i = -(k\varphi_i')' + q\varphi_i$, tenemos

$$\alpha_{ij} = (A\varphi_i, \varphi_j) = \int_0^1 \left(k \frac{d\eta_i}{dx} \frac{d\eta_j}{dx} + q\eta_i\eta_j\right) dx,$$

$$\beta_i = \int_0^1 f(x)\,\eta_i(x)\,dx. \quad (6)$$

Los cálculos nos dan

$$\frac{d\eta_i}{dx} = 0 \text{ para } x < x_{i-1},\ x > x_{i+1}$$

$$\frac{d\eta_i}{dx} \begin{cases} 1/h \text{ para } x_{i-1} < x < x_i \\ -1/h \text{ para } x_i < x < x_{i+1} \end{cases}$$

De aquí y de (6) se ve que la matriz $[a_{ij}]$ es tridiagonal, puesto que son diferentes de cero sólo aquellos α_{ij}, para los cuales $j = i - 1, i, i + 1$. Por esto, para y_i se obtiene un sistema

$$\alpha_{i,\,i-1} y_{i-1} + \alpha_{i,i} y_i + \alpha_{i,i+1} y_{i+1} - \beta_i = 0.$$

Introduciendo las designaciones

$$a_i = -h\alpha_{i,i-1} + h^2 \alpha_i = h\alpha_{i,i} + h(\alpha_{i,\,i-1} + \alpha_{i,i+1}), \quad \beta_i = -h^2\varphi_i$$

y observando que $\alpha_{i+1,\,i} = \alpha_{i,\,i+1}$, obtenemos un esquema

$$a_i y_{i-1} - (a_i + a_{i+1} + h^2 d_i)\, y_i + a_{i+1} y_{i+1} + h^2 \varphi_i = 0,$$

o bien

$$(a y_{\bar{x}})_x - dy + \varphi = 0, \quad (7)$$

donde

$$a_i = \int_{-1}^{0} k(x_i + sh)\,ds + h^2 \int_{-1}^{0} q(x_i + sh)\, s(1+s)\,ds,$$

$$d_i = \int_{-1}^{0} q(x_i + sh)(1+s)\,ds + \int_0^1 q(x_i + \text{sh})(1-s)\,ds,$$

$$\varphi_i = \int_{-1}^{0} f(x_i + sh)(1+s)\,ds + \int_0^1 f(x_i + sh)(1-s)\,ds.$$

Este es el esquema de segundo orden de aproximación.

En el método de Bubnov — Galerkin la solución u_n se busca también en la forma (6), más los coeficientes y_i se hallan de la condición de ortogonalidad del residuo $Au_n - f$ respecto de las funciones básicas $\varphi_i(x)$:

$$(Au_n - f, \varphi_i) = 0, \qquad i = 1, 2, \ldots, n, \tag{8}$$

con la particularidad de que no se requiere que el operador A sea autoconjugado. Para el problema (4) eligimos de nuevo las mismas funciones básicas. Al sustituir (6) en (8), obtendremos un sistema de ecuaciones para y_i. Calculando α_{ij} y β_i, llegamos al mismo esquema (7) que se ha obtenido por el método de Ritz.

Con la elección indicada de las funciones coordenadas $\varphi_i(x) = \eta\left(\frac{x - x_i}{h}\right)$ los métodos de Ritz y de Bubnov — Galerkin coinciden con el método de elementos finitos.

Capítulo V

Problema de Cauchy para las ecuaciones diferenciales ordinarias

En este capítulo examinaremos los esquemas de diferencias destinados para resolver ecuaciones diferenciales ordinarias (no lineales, en el caso general) de primer orden con datos iniciales (problema de Cauchy). La resolución de las ecuaciones mencionadas representa un dominio clásico de aplicación de los métodos numéricos. Existen varios métodos de diferencias, una parte de los cuales se ha elaborado en la época precedente a la invención de los ordenadores y, no obstante, resultó ser aplicable también para las máquinas electrónicas modernas. Nos limitaremos a una exposición breve de los esquemas de diferencias principales que son de amplio uso en la práctica y para los cuales se tienen los programas estándar correspondientes.

§ 1. Métodos de Runge—Kutta

1. Problema de Cauchy para una ecuación de primer orden. Supongamos que se pide hallar una función $u = u(t)$, continua para $0 \leqslant t \leqslant T$, que satisfaga la ecuación diferencial para $t > 0$ y la condición inicial para $t = 0$:

$$\frac{du}{dt} = f(t, u(t)), \quad 0 < t \leqslant T, \quad u(0) = u_0, \qquad (1)$$

donde $f(t, u)$ es la función continua prefijada de dos argumentos.

Si la función $f(t, u)$ está definida en un rectángulo $D = \{0 \leqslant t \leqslant T, |u - u_0| \leqslant U\}$ y satisface en el dominio D según la variable u la condición de Lipschitz:

$$|f(t, u_1) - f(t, u_2)| \leqslant K|u_1 - u_2|$$
$$\text{para cualesquiera } (t, u_1), (t, u_2) \in D, \qquad (2)$$

donde $K = \text{const} > 0$, entonces el problema (1) tiene una solución única.

Para demostrar esta afirmación la ecuación (1) se integra de 0 a t:

$$u(t) = u_0 + \int_0^t f(s, u(s))\, ds, \tag{3}$$

y la ecuación integral obtenida se resuelve por el método de aproximaciones sucesivas (método de Picard):

$$u_{n+1}(t) = u_0 + \int_0^t f(s, u_n(s))\, ds, \tag{4}$$

donde n es el número de la aproximación (iteración). El método de Picard converge y determina la única solución de la ecuación (3) o del problema de Cauchy (1).

Este método permite hallar la solución aproximada del problema (1), si en (4) sustituimos la integral por una fórmula de cuadratura cualquiera. Sin embargo, el volumen de los cálculos para el algoritmo obtenido es bastante grande, puesto que para cada iteración (con t fijo) debe calcularse una integral.

Para la resolución aproximada del problema (1) se emplea a veces, un método analítico basado en la idea de desarrollo de la solución del problema de Cauchy (1) en una serie de Taylor. La solución aproximada $u_n(t)$ se busca en la forma

$$u_n(t) = \sum_{k=1}^{n} \frac{t^k}{k!} u^{(k)}(0) + u_0, \quad 0 \leqslant t \leqslant T, \tag{5}$$

donde $u^{(1)}(0) = \frac{du}{dt}(0) = f(0, u_0)$, y los valores de las derivadas $u^{(k)}(0)$ $(k \geqslant 2)$ se hallan mediante la diferenciación sucesiva de la ecuación (1)

$$u^{(2)}(0) = u''(0) = \frac{d}{dt} f(t, u)\Big|_{t=0} = f_t(0, u_0) + \\ + f(0, u_0) f_u(0, u_0).$$

$$u^{(3)}(0)=u'''(0)=\frac{d^2}{dt^2}f(t,\ u)\Big|_{t=0}=$$
$$=f_{t^2}(0,\ u_0)+2f_{ut}(0,\ u_0)f(0,\ u_0)+$$
$$+f_{u^2}(0,\ u_0)u''(0),\ \ldots,$$
$$f_t=\frac{\partial f}{\partial t},\quad f_u=\frac{\partial f}{\partial u},\quad f_{ut}=\frac{\partial^2 f}{\partial u\, dt},\quad \text{etc.}$$

Para t pequeños el método de series (5) puede asegurar buena aproximación hacia la solución exacta $u(t)$ si n son no muy grandes. Aquí el volumen de los cálculos depende no sólo de la exactitud $\varepsilon>0$ ($|u(t)-u_n(t)|<\varepsilon$) y de $n=n(\varepsilon)$, sino también del tipo de la función $f(t,\ u)$, puesto que la determinación de las derivadas $u^{(k)}(t)$ puede resultar muy engorrosa.

En lo sucesivo se supondrá siempre que la función $f(t,\ u)$ es bastante suave, es decir, tiene tantas derivadas (respecto de t y de u) cuantas sean necesarias en el transcurso de la exposición.

Antes de pasar a la exposición de los esquemas de diferencias para el problema (1), detendrémonos en la cuestión de estabilidad de la solución del problema (1). ¿Cómo variará la solución del problema (1) si cambian las condiciones iniciales? Sea $\tilde{u}(t)$ la solución de la ecuación (1) con las condiciones iniciales $u(0)=\tilde{u}_0$. Para el error $z(t)=\tilde{u}(t)-u(t)$ obtenemos una ecuación

$$\frac{dz}{dt}+\alpha(t)z,\quad 0<t\leqslant T,\quad z(0)=z_0=\tilde{u}_0-u_0, \tag{6}$$

donde $\alpha(t)=[f(t,\ \tilde{u})-f(t,\ u)-f(t,\ u)]/z=f_u(t,\ u+\theta z)$, $0\leqslant\theta\leqslant 1$.

Como solución de (6) interviene la función

$$z(t)=z(0)\exp\left\{\int_0^t \alpha(s)\,ds\right\}.$$

Si $f_u\leqslant 0$ para cualesquiera $t,\ u$, entonces

$|z(t)|\leqslant|z(0)$, o bien $|\tilde{u}(t)-u(t)|\leqslant|\tilde{u}_0-u_0|$

para todo $t\in[0,\ T]$,

es decir, la solución del **problema** (1) es *estable* **respecto** de los datos iniciales (el error en los datos iniciales no crece). El problema (1) es estable también respecto del segundo miembro:

$$|\tilde{u}(t)-u(t)|\leqslant|\tilde{u}_0-u_0|+\varepsilon T \text{ para } 0\leqslant t\leqslant T,$$
$$\text{si } f_u\leqslant 0,$$

donde $\tilde{u}(t)$ es la solución del problema (1) con el segundo miembro

$$\tilde{f}=f(t,\tilde{u})+\delta f,\ |\delta f|\leqslant\varepsilon,\ \varepsilon=\text{const}>0.$$

La solución del problema (6) para $t\to\infty$ se comporta igual que la solución de una ecuación lineal

$$\frac{dz}{dt}+\lambda z=0,\quad 0<t\leqslant T,\quad z(0)=z_0,$$

que puede considerarse en el estudio de la estabilidad como la ecuación modelo.

2. Esquema de diferencias de Euler. Introduzcamos en el segmento de integración $0\leqslant t\leqslant T$ una red $\omega_\tau=\{t_n=n\tau,\ n=0,\ 1,\ \ldots\}$. Denotaremos con $y_n=y(t_n)$ una función reticular. El método numérico más simple para resolver la ecuación (1) está representado por *el esquema de diferencias de Euler*:

$$\frac{y_{n+1}-y_n}{\tau}=f(t_n,\ y_n),\quad n=0,\ 1,\ \ldots,\ y_0=u_0. \tag{7}$$

Los valores de $y_n=y(t_n)$ se determinan sucesivamente a partir de $y_0=u_0$ según una fórmula explícita

$$y_{n+1}=y_n+\tau f(t_n,\ y_n),\quad n=0,\ 1,\ \ldots,\ y_0=u_0.$$

En lugar de $u=u(t)$ encontramos una función reticular $y_n=y(t_n)$ que es la solución aproximada del problema (1).

Una función reticular

$$z_n=y_n-u(t_n)$$

es el *error* del esquema de diferencias. Escribamos la ecuación para z_n. Con este fin sustituyamos $y_n=z_n+u_n$ en

(7) y tomemos en consideración que

$$y_{n+1} - y_n = (z_{n+1} - z_n) + (u_{n+1} - u_n),$$
$$f(t_n, y_n) = f(t_n, u_n) + [f(t_n, u_n + z_n) - f(t_n, u_n)] =$$
$$= f(t_n, u_n) = \alpha_n z_n,$$

donde

$$\alpha_n = f_u(t_n, u_n + \theta z_n),\ 0 \leqslant \theta \leqslant 1.$$

Como resultado, obtenemos para z_n un problema

$$\frac{z_{n+1} - z_n}{\tau} = \alpha_n z_n + \psi_n, \quad n = 0, 1, \ldots, z_0 = 0, \qquad (8)$$

donde ψ_n es el residuo o error de la aproximación del esquema (7) en la solución $u = u(t)$ del problema (1), que es igual a

$$\psi_n = f(t_n, u_n) - \frac{u_{n+1} - u_n}{\tau}. \qquad (9)$$

Estimemos ψ_n para $\tau \to 0$. Para ello sustituyamos

$$u_{n+1} = u_n + \tau \dot{u}_n + \frac{\tau^2}{2} \ddot{u}_n + \ldots \left(\dot{u} = \frac{du}{dt}\right)$$

en (9), y, teniendo en cuenta que, de acuerdo con (1), $\dot{u}_n = f(t_n, u_n)$, obtendremos: $\psi_n = O(\tau)$, o bien $\| \psi \|_C = \max\limits_{0 \leqslant t_n \leqslant T} |\psi_n| = O(\tau)$. Esto es testimonio de que el esquema de Euler tiene *primer orden de aproximación*.

Mostremos que el esquema de Euler converge, es decir, $\| z_n \|_C = \| y_n - u_n \|_C \to 0$ para $\tau \to 0$, y tiene *primer orden de exactitud*, es decir,

$$\| z \|_C = \max_{0 \leqslant t_n \leqslant T} |z_n| = O(\tau).$$

La demostración se aduce bajo el supuesto de que

$$-K \leqslant f_u(t, u) \leqslant 0, \qquad \tau \leqslant 2/K. \qquad (10)$$

De (8) determinamos

$$z_{n+1} = (1 + \tau\alpha_n) z_n + \tau\psi_n,$$
$$|z_{n+1}| \leqslant |1 + \alpha\tau_n| \, |z_n| + \tau |\psi_n| \leqslant |z_n| + \tau |\psi_n|,$$

puesto que $|1 + \tau\alpha_n| \leqslant 1$ conforme a (10). De aquí se deduce que

$$|z_{n+1}| \leqslant |z_0| + \sum_{s=0}^{n} \tau |\psi_s| = \sum_{s=0}^{n} \tau |\psi_s|, \qquad (11)$$

es decir, $\| z \|_C = O(\tau)$.

Si la condición (10) no se cumple, pero $|f_u| \leqslant K$, entonces en lugar de (11) obtenemos $|z_{n+1}| \leqslant Te^{KT} \|\psi\|_C$, y la afirmación $|z|_C = O(\tau)$ queda en vigor.

3. Aumento del orden de exactitud. El método de Euler es muy sencillo, mas no es de elevada exactitud. Se puede aumentar el orden de exactitud de la solución numérica respecto de τ sin complicar el algoritmo. Existe el *método de Runge* para elevar la exactitud cuya idea consiste en lo siguiente. Supongamos que la solución $u = u(t)$ es suficientemente suave y tiene lugar el desarrollo siguiente del error $z_n = y_n - u_n$ en potencias de τ:

$$y_n = u_n + \alpha(t)\tau + \beta(t)\tau^2 + \ldots, \qquad (12)$$

donde $\alpha(t)$ y $\beta(t)$ son unas funciones que no dependen de τ.

Elijamos dos redes de pasos τ_1 y τ_2 que tienen nodos comunes (por ejemplo, $\tau_1 = \tau$, $\tau_2 = \tau/2$), resolvamos en cada red el problema (7) y encontremos $y^{(1)}(t_{n_1})$ e $y^{(2)}(t_{n_2})$, respectivamente. Tomemos un nodo, común para las dos redes, $t_n^* = t_{n1} = t_{n2}$, y escribamos (12) para $n = n^*$:

$$y^{(1)}(t_{n^*}) = u(t_{n^*}) + \alpha(t_{n^*})\tau_1 + O(\tau_1^2),$$
$$y^{(2)}(t_{n^*}) = u(t_{n^*}) + \alpha(t_{n^*})\tau_2 + O(\tau_2^2).$$

Formemos una combinación lineal con el parámetro σ:

$$\tilde{y}(t_{n^*}) = \sigma y^{(1)}(t_{n^*}) + (1 - \sigma) y^{(2)}(t_{n^*}) =$$
$$= u(t_{n^*}) + [\sigma\tau_1 + (1 - \sigma)\tau_2]\alpha(t_{n^*}) + O(\tau_1^2 + \tau_2^2).$$

Eligiendo σ de la condición $\sigma\tau_1 + (1 - \sigma)\tau_2 = 0$, es decir suponiendo $\sigma = \tau_2/(\tau_2 - \tau_1)$, obtenemos

$$\tilde{y}(t_{n^*}) = u(t_{n^*}) + O(\tau^2), \quad \tau = \text{máx}(\tau_1, \tau_2).$$

La función reticular $\tilde{y}$ aproxima la solución $u = u(t)$ con el segundo orden de exactitud respecto de τ. De este modo, he-

mos elevado la exactitud del método de Euler realizando dos cálculos en las redes de pasos τ_1 y τ_2. Este procedimiento puede ser continuado teniendo presente (12). Al realizar los cálculos según el esquema (7) en tres redes de pasos τ_1, τ_2, τ_3, hallaremos la solución del problema (1) con el tercer orden de exactitud en los nodos comunes para las tres redes elegidas.

4. Esquemas de Runge-Kutta. El orden de exactitud puede ser aumentado complicando el esquema de diferencias. Son de amplio uso en la práctica los *esquemas de Runge-Kutta* del segundo y cuarto órdenes de exactitud.

El cálculo por el esquema de Runge—Kutta del segundo orden de exactitud se realiza en dos etapas. En la primera etapa se halla el valor intermedio de $\overline{y}_n$ según el esquema de Euler de paso $\alpha\tau$:

$$\overline{y}_n = y_n + \alpha\tau f(t_n, y_n);$$

en la segunda etapa se determina el valor de y_{n+1} por la fórmula

$$y_{n+1} = y_n + \tau(1-\sigma) f(t_n, y_n) + \sigma\tau f(t_n + \alpha\tau, \overline{y}_n),$$

donde $\alpha > 0$, $\sigma > 0$ son los parámetros. Al eliminar $\overline{y}_n$, obtendremos para y_{n+1} un esquema

$$\frac{y_{n+1} - y_n}{\tau} = (1-\sigma) f(t_n, y_n) + \sigma f(t_n + \alpha\tau, y_n + \alpha\tau f(t_n, y_n)). \quad (13)$$

El orden de exactitud del esquema depende de los parámetros α, τ.

Hallemos una expresión para el residuo o error de aproximación del esquema (13). Con este fin, por analogía con el p. 2, traslademos $(y_{n+1} - y_n)/\tau$ en el segundo miembro y sustituyamos u_n, u_{n+1} en lugar de y_n, y_{n+1}. De resultas, obtendremos la siguiente expresión para el residuo:

$$\psi_n = (1-\sigma) f(t_n, u_n) + \sigma f(t_n + \alpha\tau, u_n + \alpha\tau f(t_n, u_n)) - (u_{n+1} - u_n)/\tau. \quad (13')$$

Recurriendo al desarrollo por la fórmula de Taylor, obtenemos

$$\psi_n = \tau(\sigma\alpha - 1/2) u''_n + O(\tau^2).$$

De aquí se ve que el esquema (13) tiene segundo orden de aproximación $\psi_n = O(\tau^2)$, si se cumple la condición

$$\sigma\alpha = 1/2. \qquad (14)$$

De este modo, existe una familia (de un solo parámetro) de esquemas (13), (14) de segundo orden de aproximación.

Veamos los casos particulares:

1) $\sigma = 1$, $\alpha = 1/2$:

$$\frac{\bar{y}_n - y_n}{\tau/2} = f(t_n y_n), \quad \frac{y_{n+1} - y_n}{\tau} = f\left(t_n + \frac{\tau}{2}, \bar{y}_n\right). \qquad (15)$$

Este es el conocido esquema *predictor—corrector*, o bien *cálculo—recálculo*. Puede ser escrito de otra forma:

$$\bar{y}_n = y_n + \frac{\tau}{2} f(t_n, y_n), \quad y_{n+1} = y_n + \tau f\left(t_n + \frac{\tau}{2}, \bar{y}_n\right),$$

o, al eliminar $\bar{y}_n$, en la forma

$$(y_{n+1} - y_n)/\tau = f\left[t_n + \frac{\tau}{2}, y_n + \frac{\tau}{2} f(t_n, y_n)\right]. \qquad (15')$$

2) $\sigma = 1/2$, $\alpha = 1$:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{1}{2}[f(t_n, y_n) + f(t_{n+1}, y_n + \tau f(t_n, y_n))]. \qquad (16)$$

Este esquema también puede considerarse como un esquema predictor—corrector: al principio, el esquema de Euler de paso τ (*predictor*):

$$\bar{y}_n = y_n + \tau f(t_n y_n);$$

después, el esquema con una semisuma (*corrector*):

$$(y_{n+1} - y_n)/\tau - {}^1/_2 [f(t_n, y_n) + f(t_{n+1}, \bar{y}_n)].$$

La idea del método predictor—corrector se usa con frecuencia al escribir esquemas de diferencias para las ecuaciones de la física matemática con derivadas parciales.

He aquí las fórmulas para el esquema de Runge—Kutta del cuarto orden de exactitud:

$$\frac{y_{n+1} - y_n}{\tau} = \frac{1}{6}[k_1(y_n) + 2k_2(y_n) + 2k_3(y_n) + k_4(y_n)],$$

$$n = 0, 1, \ldots, y_0 = u_0, \qquad (17)$$

donde k_1, k_2, k_3, k_4 son las correcciones que se calculan según las fórmulas

$$k_1 = f(t_n, y_n),\ k_2 = f(t_n + \tau/2,\ y_n + \tau k_1/2),$$
$$k_3 = f(t_n + \tau/2,\ y_n + \tau_{k_2}/2),\ k_4 = f(t_n + \tau,\ y_n + \tau k_3). \tag{18}$$

Determinando y_{n+1} según el y_n dado se debe cuatro veces calcular el segundo miembro.

Demos a conocer el método de los cálculos según este esquema. Para $n = 0$ sabemos $y_0 = u_0$. Podemos calcular sucesivamente k_1, k_2, k_3, k_4, y hallar

$$y_1 = y_0 + \frac{1}{6}\tau(k_1(y_0) + 2k_2(y_0) + 2k_3(y_0) + k_4(y_0)),$$

después de lo cual los cálculos se realizan para $n = 1, 2, \ldots$ Para el residuo obtenemos una expresión

$$\psi_n = \frac{1}{6}[k_1(u_n) + 2k_2(u_n) + 2k_3(u_n) + k_4(u_n)] - \\ - \frac{u_{n+1} - u_n}{\tau}, \tag{19}$$

donde $k_i(u_n)$ ($i = 1, 2, 3, 4$) se determinan según las fórmulas (18), en las cuales y_n se ha sustituido por u_n.

Al desarrollar u_{n+1}, $k_2(u_n)$, $k_3(u_n)$, $k_4(u_n)$ en el entorno de $t = t_0$, nos convencemos de que $\psi_n = O(\tau^4)$, es decir, el esquema (7), (18) tiene el cuarto orden de aproximación, si $u = u(t)$ tiene cuatro derivadas continuas.

Todos los métodos de Runge—Kutta son *explícitos* (para determinar y_{n+1} se debe realizar los cálculos según las fórmulas explícitas) y de *un paso* (para determinar y_{n+1} se debe hacer un paso en la red desde t_n hasta t_{n+1}).

5. Estabilidad de los esquemas de diferencias. En el p. 1 se ha considerado una propiedad importante de la ecuación diferencial (1), a saber, la estabilidad (respecto de los datos iniciales y del segundo miembro). Para estudiar la estabilidad respecto de los datos iniciales de la ecuación no lineal (1) analizaremos una ecuación modelo

$$\frac{du}{dt} + \lambda u = 0,\quad \lambda = \text{const} > 0,\quad t > 0,\quad u(0) = u_0. \tag{20}$$

Su solución $u(t) = u_0 e^{-\lambda t}$ decrece para $\lambda > 0$, y

$$|u(t)| \leqslant |u_0| \text{ cuando } \lambda \geqslant 0 \text{ para todo } t \geqslant 0, \quad (21)$$

es decir, la ecuación (20) es *estable para* $\lambda \geqslant 0$, lo que corresponde a la condición $f_u \leqslant 0$.

Se introduce una exigencia natural: para los esquemas de diferencias que aproximan las ecuaciones modelo ha de cumplirse un análogo de la desigualdad (21):

$$|y_n| \leqslant |y_0| \text{ para cualesquiera } n = 1, 2, \ldots \quad (22)$$

Veremos más abajo que esto no siempre se cumple.

Veamos una serie de ejemplos.

1) ESQUEMA EXPLICITO DE EULER:

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_n = 0, \qquad y_{n+1} = (1 - \tau\lambda)\, y_n. \quad (23)$$

De aquí se ve que la condición

$$|y_{n+1}| \leqslant |y_n| \leqslant \ldots \leqslant |y_0| \quad (24)$$

queda cumplida para $|1 - \tau\lambda| \leqslant 1$, o bien $-1 \leqslant 1 - \tau\lambda \leqslant 1$, es decir, para

$$\tau\lambda \leqslant 2. \quad (25)$$

Si, por ejemplo, $\tau\lambda \leqslant 3$, entonces

$$|y_{n+1}| = |\tau\lambda - 1|\,|y_n| \geqslant 2\,|y_n| \geqslant \ldots \geqslant 2^{n+1}\,|y_0|,$$
$$|y_n| \geqslant 2^n\,|y_0| \to \infty \text{ cuando } n \to \infty.$$

El esquema es inestable, la condición (24) no se cumple. De este modo, el esquema de Euler (23) es *convencionalmente estable* para $\tau \leqslant 2/\lambda$, $\lambda > 0$.

2) Esquema implícito de Euler:

$$\frac{y_{n+1} - y_n}{\tau} + \lambda y_{n+1} = 0 \qquad y_{n+1} = \frac{1}{1 + \tau\lambda}\, y_n. \quad (26)$$

Por cuanto $1/(1 + \tau\lambda) \leqslant 1$ para cualesquiera $\tau\lambda \geqslant 0$, entonces el esquema es *absolutamente estable*:

$$|y_n| \leqslant |y_0| \text{ para cualesquiera } \tau \text{ y } \lambda \geqslant 0, \; n = 0, 1, 2, \ldots \quad (27)$$

3) ESQUEMA CON PESOS

$$\frac{y_{n+1}-y_n}{\tau}+\lambda\left(\sigma y_{n+1}+(1-\sigma)\,y_n\right)=0,\quad y_{n+1}=qy_n. \qquad (28)$$

El esquema es estable para

$$|\,q\,|\leqslant 1,\qquad q=\frac{1-(1-\sigma)\,\tau\lambda}{1+\sigma\tau\lambda}.$$

Vemos que $|\,q\,|\leqslant 1$, si $-1-\sigma\tau\lambda\leqslant 1-(1-\sigma)\,\tau\lambda\leqslant$ $\leqslant 1+\sigma\tau\lambda$, o bien $1+\tau\,(\sigma-1/2)\,\lambda\geqslant 0$, de modo que $1+\sigma\tau\lambda\geqslant\tau\lambda/2>0$. De este modo el *esquema con pesos es absolutamente (para todo τ) estable* para $\sigma>1/2$, y *condicionalmente estable* para $\sigma<1/2$, siempre que $\tau\leqslant$ $\leqslant 1/((1/2-\sigma)\,\lambda)$.

4. ESQUEMA DE RUNGE-KUTTA DE SEGUNDO ORDEN. Al sustituir en la fórmula (13) $f=-\lambda y$, obtenemos

$$y_{n+1}=qy_n,\qquad q=1-\tau\lambda+\frac{1}{2}\tau^2\lambda^2. \qquad (29)$$

El esquema es estable, $|\,y_n\,|\leqslant|\,y_0\,|$, si $|\,q\,|=1-\tau\lambda+$ $+\frac{1}{2}\tau^2\lambda^2\leqslant 1$, lo que tiene lugar cuando

$$\tau\lambda\leqslant 2. \qquad (25)$$

El esquema de Runge—Kutta de segundo orden es estable bajo la misma condición que el esquema explícito de Euler.

5) ESQUEMA DE RUNGE–KUTTA DE CUARTO ORDEN. Sustituyendo $f=-\lambda y$ en (17), (18), obtenemos

$$y_{n+1}=qy_n,$$
$$q=1-\tau\lambda+\frac{1}{2}\tau^2\lambda^2-\frac{1}{6}\tau^3\lambda^3+\frac{1}{24}\tau^4\lambda^4. \qquad (30)$$

La desigualdad $|\,q\,|\leqslant 1$ se cumple para $\tau\lambda\leqslant 2{,}78$, es decir, la condición de estabilidad del esquema de cuarto orden es *un poco más débil que la condición* (25) para el esquema de segundo orden.

Estos ejemplos muestran que los esquemas explícitos de *un paso son condicionalmente estables*, y entre los esquemas implícitos se tienen absolutamente estables (por ejemplo (28) cuando $\sigma\geqslant 1/2$). Si $\lambda>0$ es grande, el paso τ, en virtud de (25), debe elegirse para los esquemas explícitos lo suficientemente pequeño.

6. Sobre la convergencia y la exactitud. El esquema de Runge—Kutta para una ecuación no homogénea

$$\frac{du}{dt}+\lambda u=f(t),\quad t>0,\quad u(0)=u_0 \tag{31}$$

tiene por expresión

$$y_{n+1}=qy_n+\tau\varphi_n,\quad q=q(\tau\lambda), \tag{32}$$

donde las expresiones para q y φ_n dependen del orden del esquema. Así, para el esquema de segundo orden tenemos

$$q=1-\tau\lambda+\frac{1}{2}\tau^2\lambda^2,$$

$$\varphi_n=(1-\sigma)f(t_n)+\sigma f(t_n+\alpha\tau),\quad \alpha\sigma=\frac{1}{2}.$$

Para el error $z_n=y_n-u_n$ obtenemos

$$\frac{z_{n+1}-z_n}{\tau}+\left(\lambda-\frac{\lambda^2\tau}{2}\right)z_n=\psi_n$$

o bien

$$z_{n+1}=qz_n+\tau\psi_n,\ n=0,\ 1,\ 2,\ \ldots,\ z_0=0,$$

donde ψ_n es el residuo igual a

$$\psi_n=\varphi_n-(u_{n+1}-u_n)/\tau=O(\tau^2).$$

En virtud de la condición de estabilidad (25), $|q|\leqslant 1$ y

$$|z_{n+1}|\leqslant|z_n|+\tau|\psi_n|\leqslant\sum_{k=0}^{n}\tau|\psi_k|, \tag{33}$$

de donde precisamente proviene que el esquema (32) converge y tiene el segundo orden de exactitud (converge con la velocidad $O(\tau^2)$, o converge con el segundo orden):

$$\|z\|_C=O(\tau^2).$$

De este modo, si un esquema es estable y aproxima la ecuación (1), es convergente. Esta afirmación demostrada para el problema modelo tiene un carácter general y es verídica para cualquiera de los esquemas de segundo orden.

De modo análogo se demuestra la convergencia con la velocidad $O(\tau^2)$ del esquema de Runge—Kutta (13) a con-

dición de que $f_u \leqslant 0$. En este caso, para $z_n = y_n - u_n$ con $\sigma\alpha = {}^1/_2$ obtenemos un problema

$$\frac{z_{n+1}-z_n}{\tau} = \beta_n \left(1 + \frac{1}{2}\tau\gamma_n\right) z_n + \tau\psi_n, \qquad (34)$$

donde $\beta_n = f_u(t_n, u_n + \theta_1 z_n)$, $\gamma_n = f_u(t_n (+ \tau/2, u_n + \theta_2 z_n)$ $(0 \leqslant \theta_i \leqslant 1, i = 1, 2)$, y ψ se determina por la fórmula (13'). Reescribamos (34) en la forma

$$z_n = q_n z_n + \tau\psi_n, \quad q_n = 1 + \tau\beta_n (1 + \tau\gamma_n/2).$$

La condición de estabilidad $|q_n| \leqslant 1$ ó $-1 \leqslant q_n \leqslant 1$ se cumple, si $2 - \tau|\beta_n| + {}^1/_2\tau^2 |\beta_n| |\gamma_n| \geqslant 0$,
${}^1/_2\tau |\beta_n| |\gamma_n| \leqslant |\beta_n|$, o bien $\tau |\gamma_n| \leqslant 2$. La primera desigualdad queda cumplida también para $\tau |\beta_n| \leqslant 2$, y, por consiguiente, es suficiente que sea

$$\tau K \leqslant 2, \qquad (35)$$

siempre que $f_u \leqslant 0$, $|f_u| \leqslant K$, $(t, u) \in D$. La condición (35) es análoga a (25) y asegura el cumplimiento de la estimación (33), de la cual se deduce precisamente la convergencia del esquema (13) con el segundo orden $\| z \|_C = O(\tau^2)$.

§ 2. Esquemas de varios pasos. Métodos de Adams

1. Esquemas de varios pasos. En el § 1 fueron considerados los métodos de Runge—Kutta para la resolución numérica del problema de Cauchy

$$\frac{du}{dt} = f(t, u), \quad 0 < t \leqslant T, \quad u(0) = u_0. \qquad (1)$$

Estos son los *métodos de un paso*: al determinar el nuevo valor de y_{n+1} se usa sólo el valor de y_n. Para determinar el valor aproximado de y_n pueden analizarse, en el caso general, los *esquemas de diferencias de m pasos* $(m \geqslant 1)$, es decir, las ecuaciones del tipo

$$\sum_{k=0}^{m} \frac{a_k}{\tau} y_{n-k} = \sum_{k=0}^{m} b_k f_{n-k}, \quad n = m, m+1, \ldots, \qquad (2)$$

donde a_k, b_k son ciertos coeficientes numéricos,

$$f_{n-k} = f(t_{n-k}, y_{n-k}), \quad a_0 \neq 0, \quad b_m \neq 0.$$

En particular, para $m = 1$, $b_0 = 0$, $b_1 = -a_0$, $a_1 = -a_0$, llegamos al esquema de Euler.

El esquema (2) se denomina explícito (de extrapolación), si $b_0 = 0$ y los valores de y_n se determinan en términos de los valores antecedentes de $y_{n-1}, y_{n-2}, \ldots, y_{n-m}$ según una fórmula explícita

$$y_n = \frac{1}{a_0} \sum_{k=1}^{m} (b_k \tau f_{n-k} - a_k y_{n-k}) = \frac{1}{a_0} F(y_{n-1}, y_{n-2}, \ldots, y_{n-m}).$$

Los cálculos empiezan con $n = m$. Para hallar y_m, se deben prefijar m valores iniciales $y_0, y_1, \ldots, y_{m-1}$; éstos pueden determinarse, por ejemplo, mediante el método de Runge—Kutta en el que se utiliza solamente el valor inicial de $y_0 = u_0$.

Si $b_0 \neq 0$, el esquema (2) se denomina implícito (de interpolación): al hallar y_n, se debe resolver, para cada n, una ecuación no lineal

$$a_0 y_n - b_0 f(t_n, y_n) = F(y_{n-1}, y_{n-2}, \ldots, y_{n-m}). \tag{3}$$

Dicha ecuación no lineal puede resolverse, por ejemplo, mediante el método de Newton.

El error de aproximación del esquema (2) en la solución $u = u(t)$ de la ecuación (1) o el residuo se determina por la fórmula

$$\psi_n = \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \sum_{k=0}^{m} a_k u_{n-k}. \tag{4}$$

Suele decirse que el esquema (2) tiene el s-ésimo *orden de aproximación* (o simplemente que el esquema (2) tiene el s-ésimo *orden*), si

$$\| \psi \|_C = O(\tau^s), \quad \text{o bien } \| \psi \|_C \leqslant M\tau^s, \quad s > 0, \tag{5}$$

donde $M = \text{const} > 0$ no depende de τ.

Los coeficientes a_k, b_k se eligen a partir de los requisitos de aproximación y estabilidad. Sin perturbar la generalidad

de razonamientos podemos considerar que

$$\sum_{k=0}^{m} b_k = 1, \tag{6}$$

puesto que los coeficientes de la ecuación (2) están determinados con una exactitud de hasta un factor. Desarrollando ψ_n en potencias de τ y exigiendo que el residuo tenga un orden prefijado, obtenemos las condiciones para determinar a_k, b_k. Por cuanto $u = 1$ es la solución de la ecuación $u_t = f(t, u)$ para $f = 0$, de (2) se desprende que

$$\sum_{k=0}^{m} a_k = 0. \tag{7}$$

Con el objeto de construir los esquemas (2) se aplican, corrientemente, otros procedimientos en los cuales se emplean fórmulas de cuadratura y de interpolación. Así por ejemplo, integrando la ecuación diferencial (1) respecto de t dentro de los límites de t_{n-n_0} a t_n, obtenemos

$$u_n - u_{n-n_0} = \int_{t_{n-n_0}}^{t_n} f(t, u(t))\, dt. \tag{8}$$

Para obtener de aquí un esquema de diferencias se puede usar para la integral una fórmula de cuadratura cualquiera.

2. Método de Adams. Toda fórmula de cuadratura engendra el método correspondiente de resolución numérica de la ecuación diferencial ordinaria (1). En una identidad

$$u_n - u_{n-1} = \int_{t_{n-1}}^{t_n} f(t, u(t))\, dt, \tag{9}$$

correspondiente a la identidad (8) cuando $n_0 = 1$, sustituyamos la integral por una fórmula de cuadratura

$$\int_{t_{n-1}}^{t_n} f(t, u(t))\, dt \approx \tau \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}). \tag{10}$$

Teniendo presente (9) y (10), podemos escribir el *esquema de diferencias de Adams*:

$$\frac{y_n - y_{n-1}}{\tau} = \sum_{k=0}^{m} b_k f(t_{n-k}, y_{n-k}). \qquad (11)$$

Dicho esquema puede ser obtenido de (2), si ponemos $a_k = 0$ para $k = 2, 3, \ldots, m$, y $a_0 = 1$, $a_1 = -1$.

La fórmula de cuadratura (10), en cuya base está construido el esquema de Adams, contiene los nodos de las redes que no pertenecen al intervalo de integración $t_{n-1} \leqslant t \leqslant t_n$. Habitualmente se utiliza la exigencia de que la fórmula de cuadratura sea exacta para un polinomio de grado m. Con ello se elige un polinomio de interpolación con los nodos t_n, $t_{n-1}, \ldots, t_{n-m}$.

Con tal construcción del esquema su error de aproximación coincide con el error de la fórmula de cuadratura. Efectivamente, el residuo para el esquema (11) es

$$\psi_n = \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}) - \frac{u_n - u_{n-1}}{\tau}.$$

Sustituyendo aquí de (9) la expresión

$$\frac{u_n - u_{n-1}}{\tau} = \frac{1}{\tau} \int_{t_{n-1}}^{t_n} f(t, u(t))\, dt,$$

obtenemos la fórmula para el residuo:

$$\psi_n = \sum_{k=0}^{m} b_k f(t_{n-k}, u_{n-k}) - \frac{1}{\tau} \int_{t_{n-1}}^{t_n} f(t, u(t))\, dt. \qquad (12)$$

3. **Esquemas explícitos e implícitos.** Si $b_0 = 0$, el esquema (11) será explícito y

$$y_n = y_{n-1} + \tau \sum_{k=1}^{m} b_k f_{n-k}. \qquad (13)$$

De ejemplo más simple del esquema explícito de Adams sirve el de Euler

$$y_n - y_{n-1} = \tau f_{n-1} \quad \text{para } m = 1,\ b_0 = 0,\ b_1 = 1. \qquad (14)$$

Al poner en (11) $m = 1$, $b_0 = 1$, $b_1 = 0$, obtendremos el esquema de Adams implícito

$$\frac{y_n - y_{n-1}}{\tau} = f_n, \quad \text{o bien} \quad y_n - \tau f(t_n, y_n) = y_{n-1}. \qquad (15)$$

El esquema implícito *simétrico* de un paso ($m = 1$)

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{2}[f(t_n, y_n) + f(t_{n-1}, y_{n-1})] \qquad (16)$$

corresponde a los valores $m = 1$, $b_0 = b_1 = 1/2$ y tiene el segundo orden de aproximación: $\psi_n = O(\tau^2)$. Para determinar y_n se debe resolver (con cada n) una ecuación no lineal $y_n = {}^1/_2\tau f(t_n, y_n) = F_{n-1}$, donde $F_{n-1} = y_{n-1} + {}^1/_2\tau f(t_{n-1}, y_{n-1})$.

Veamos ahora los esquemas de Adams de dos pasos que corresponden a $m = 2$. El esquema explícito de dos pasos ($m = 2$) tiene por expresión

$$\frac{y_n - y_{n-1}}{\tau} = \frac{3}{2} f_{n-1} - \frac{1}{2} f_{n-2},$$

$$m = 2, \quad b_0 = 0, \quad b_1 = \frac{3}{2}, \quad b_2 = -\frac{1}{2}. \qquad (17)$$

El esquema es de segundo orden de aproximación:

$$\psi_n = \frac{3}{2} f(t_{n-1}, u_{n-1}) - \frac{1}{2} f(t_{n-2}, u_{n-2}) - \frac{u_n - u_{n-1}}{\tau} = O(\tau^2).$$

Investiguemos la estabilidad del esquema modelo correspondiente

$$\frac{y_n - y_{n-1}}{\tau} + \lambda\left(\frac{3}{2} y_{n-1} - \frac{1}{2} y_{n-2}\right) = 0. \qquad (18)$$

Al sustituir aquí $y_n = q^n$, obtendremos

$$q^2 - \left(1 - \frac{3}{2}\mu\right) q - \frac{1}{2}\mu = 0, \quad \mu = \lambda\tau. \qquad (19)$$

Por cuanto $D = 1 - \mu + \frac{9}{4}\mu^2 > 0$ para μ cualquiera, las raíces q_1 y q_2 serán reales y distintas. La estabilidad significa que $|q_1| \leqslant 1$ y $|q_2| \leqslant 1$. Hagamos uso de la siguien-

te propiedad que se comprueba inmediatamente: las raíces de una ecuación cuadrática $q^2 + bq + c = 0$ no superan en módulo la unidad:

$$|q_{1,2}| \leqslant 1, \text{ si } |b| \leqslant 1 + c, \ c \leqslant 1. \tag{20}$$

Para la ecuación (19) tenemos $b = 3\mu/2 - 1$, $c = -\mu/2$, y la condición $|3\mu/2 - 1| \leqslant 1 - \mu/2$ queda cumplida para $\mu \leqslant 1$, ó

$$\tau\lambda \leqslant 1,$$

es decir, el esquema (18) es condicionalmente estable (el paso τ debe ser dos veces menor que el paso admisible en el esquema de Euler).

Escribamos un esquema implícito de Adams de dos pasos ($m = 2$). Exigiendo que la fórmula de cuadratura (10) sea exacta para los polinomios de grados 0, 1, 2, es decir, que $F(t) = f(t, u(t)) = \{1, t, t^2\}$, encontramos los coeficientes $b_0 = 5/12$, $b_1 = 8/12$, $b_2 = -1/12$. El esquema es de la forma

$$\frac{y_n - y_{n-1}}{\tau} = \frac{1}{12}(5f_n + 8f_{n-1} - f_{n-2}). \tag{21}$$

Investiguemos la estabilidad del problema modelo

$$\frac{y_n - y_{n-1}}{\tau} + \frac{\lambda}{12}(5y_n + 8y_{n-1} - y_{n-2}) = 0. \tag{22}$$

Suponiendo $y_n = q^n$, obtendremos una ecuación característica

$$aq^2 + bq + c = 0, \ a = 1 + \frac{5}{12}\tau\lambda, \qquad b = \frac{8}{12}\tau\lambda - 1,$$
$$c = -\frac{1}{12}\tau\lambda.$$

Las condiciones (20), para las cuales $|q_{1,2}| \leqslant 1$, adquieren la forma $|b| \leqslant a + c$, $c \leqslant a$. De aquí se infiere que el esquema (22) es estable cuando $\tau\lambda \leqslant 6$.

4. Problema de Cauchy para una ecuación de segundo orden. Analicemos un problema de Cauchy:

$$\frac{d^2u}{dt^2} = f(t, u(t)), \ t > 0, \ u(0) = u_0, \quad \frac{du}{dt}(0) = u_1. \tag{23}$$

Los más usados son los *métodos de Störmer*

$$\frac{y_{n+1}-2y_n+y_{n-1}}{\tau^2}=\sum_{k=-1}^{m} b_k f(t_{n-k},\ y_{n-k}),$$

$$m \geqslant 0,\ n=1,\ 2,\ \ldots, \qquad (24)$$

$$y_0=u_0, \quad y_1=\bar{u}_1 \text{ o bien } \frac{y_1-y_0}{\tau}=\tilde{u}_1.$$

El valor de $\bar{u}_1$ (o de $\tilde{u}_1$) se elige de una manera tal que el error de aproximación $\nu = \frac{1}{\tau}[u(\tau) - u(0)] - \dot{u}(0) - \tilde{u}_1$ tenga cierto orden, por ejemplo, $\nu = O(\tau^p)$, donde p es el orden de aproximación del esquema (24). Por ejemplo, para $p = 2$ hallamos

$$u(\tau)=u(0)+\tau\dot{u}(0)+\frac{1}{2}\tau^2\ddot{u}(0)+O(\tau^3),$$

$$\nu=u_1+\frac{\tau}{2}\ddot{u}(0)-\tilde{u}_1+O(\tau^2)=\frac{\tau}{2}f(0,\ u(0))+$$

$$+O(\tau^2)-\tilde{u}_1+u_1=O(\tau^2),$$

si ponemos

$$\tilde{u}_1 = u_1 + {}^1/_2\tau f(0,\ u_0),\ \bar{u}_1 = u_0 + \tau\tilde{u}_1.$$

Si $b_{-1} = 0$, el esquema (24) será explícito, puesto que en el segundo miembro figuran sólo valores conocidos de y_n, $y_{n-1}, \ldots, y_{n-m}$. Si $b_{-1} = 0$, el esquema (24) es implícito y para determinar y_{n+1} se debe resolver la ecuación

$$y_{n+1} - b_{-1}f(t_{n+1},\ y_{n+1}) = F(y_n,\ y_{n-1},\ \ldots,\ y_{n-m},\ t_n).$$

Para obtener el esquema de diferencias (24) calculemos una integral

$$\int_{t_{n-1}}^{t_{n+1}} u''v\,dt=\int_{t_{n-1}}^{t_n} u''v\,dt+\int_{t_n}^{t_{n+1}} u''v\,dt=$$

$$=(u'v-uv')\big|_{t_{n-1}}^{t_n}+(u'v-uv')\big|_{t_{n-1}}^{t_n}+\int_{t_{n-1}}^{t_n} uv''\,dt, \qquad (25)$$

donde $v(t)$ es una función continua a trozos

$$v(t)=\begin{cases}(t-t_{n-1})/\tau & \text{para } t_{n-1}\leqslant t\leqslant t_n,\\ (t_{n+1}-t)/\tau & \text{para } t_n\leqslant t\leqslant t_{n+1}.\end{cases} \tag{26}$$

Sustituyamos (26) en (25) teniendo presente que $v''(t)=0$:

$$\int_{t_{n-1}}^{t_{n+1}} u''v\,dt=\frac{1}{\tau}(u_{n-1}-2u_n+u_{n+1}). \tag{27}$$

Luego, multiplicando la ecuación (23) por $v(t)$ y tomando en consideración (27), obtendremos una identidad

$$\frac{u_{n+1}-2u_n+u_{n-1}}{\tau}=\frac{1}{\tau}\int_{t_{n-1}}^{t_{n+1}} f(t,\ u(t))\ v(t)\,dt. \tag{28}$$

El error de aproximación del esquema (24) en la solución $u=u(t)$, o el residuo para el esquema (24) se determina mediante la fórmula

$$\psi_n=\sum_{k=-1}^{m} b_k f(t_{n-k},\ u_{n-k})-\frac{u_{n+1}-2u_n+u_{n-1}}{\tau^2},$$

la cual puede ser escrita, en virtud de la identidad (28), en la forma

$$\psi_n=\sum_{k=-1}^{m} b_k f(t_{n-k},\ u_{n-k})-\frac{1}{\tau}\int_{t_{n-1}}^{t_{n+1}} f(t,\ u(t))\,v(t)\,dt. \tag{29}$$

Al introducir una nueva variable $s=(t-t_n)/\tau$, escribamos la integral en la forma más cómoda:

$$\frac{1}{\tau}\int_{t_{n-1}}^{t_{n+1}} F(t)\,v(t)\,dt=\int_{-1}^{1} F(t_n+s\tau)\,\bar{v}(s)\,ds,$$

$$F=f(t,\ u(t)),\quad \bar{v}(s)=\begin{cases}1+s, & s<0,\\ 1-s, & s>0.\end{cases}$$

De (29) se ve que el primer sumando es una fórmula de cuadratura para la integral de la función $F(t)=f(t,\ u(t))$

con el peso $v(t) \geqslant 0$. El error de aproximación del esquema se determina completamente por el de la fórmula de cuadratura. Los métodos construidos en esta base se denominan, además ***métodos de Adams—Störmer***.

La fórmula más simple de un rectángulo da un esquema

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} = f(t_n, y_n),$$

puesto que $\frac{1}{\tau} \int\limits_{t_{n-1}}^{t_{n+1}} v(t)\, dt = 1.$

Para el problema modelo

$$\frac{d^2u}{dt^2} + \lambda u = 0, \quad t > 0, \quad u(0) = 0, \quad \frac{du}{dt}(0) = u_1$$

tenemos

$$\frac{y_{n+1} - 2y_n + y_{n-1}}{\tau^2} + \lambda y_n = 0.$$

Sustituyendo aquí $y_n = q^n$, encontramos $q^2 - 2(1 - \tau^2\lambda/2)q + 1 = 0$; $D < 0$ para $\lambda\tau^2 \leqslant 4$, $\tau \leqslant 2/\sqrt{\lambda}$; además, $|q_1| = |q_2|$ y el esquema es estable a condición de que $\tau \leqslant 2\sqrt{\lambda}$ ó $\tau\sqrt{\lambda} \leqslant 2$.

5. Sistemas de ecuaciones. Muchos métodos se extienden sin cambios algunos al problema de Cauchy para el sistema de ecuaciones

$$\frac{du}{dt} = f(t, u), \quad t > 0, \quad u(0) = u_0, \tag{30}$$

donde $u = (u^1(t), u^2(t), \ldots, u^N(t))$ es el vector buscado, y $f = (f^1, f^2, \ldots, f^N)$, el vector prefijado. Escribamos (30) en componentes

$$\frac{du^i}{dt} = f^i(t, u), \quad t > 0, \quad u^i(0) = u_0^i, \quad i = 1, 2, \ldots, N. \tag{31}$$

Sean u, v dos soluciones del problema (30) con los datos iniciales $u(0) = u_0$, $v(0) = v_0$. Para su diferencia $z =$

$= u - v$ obtendremos un sistema de ecuaciones lineales

$$\frac{dz^i}{dt} = \sum_{j=1}^{n} \alpha_{ij}(t)\, z^j,$$

donde α_{ij} es el valor de la derivada $\partial f^i/\partial u^j$ en cierto punto medio $(t, \overline{u}_j)$, $\overline{u}_j = (v^1, v^2, \ldots, v^{j-1}, u^j + \theta, z^j, u^{j+1}, \ldots$ $\ldots, u^N)$ $(0 \leqslant \theta_j \leqslant 1, j = 1, 2, \ldots, N)$. Por eso el modelo lineal del sistema de ecuaciones no lineales (30) será representado por un sistema lineal

$$\frac{du^i}{dt} + \sum_{j=1}^{N} a_{ij} u^j = f^i(t). \tag{32}$$

o, en la forma vectorial,

$$\frac{du}{dt} + Au = f(t), \qquad A = (a_{ij}). \tag{33}$$

Para que dicha ecuación sea estable respecto de los datos iniciales, es suficiente que la matriz A sea no negativa. En el párrafo que sigue se indicarán las condiciones necesarias y suficientes de estabilidad de los esquemas para los sistemas de ecuaciones lineales (33).

En la práctica nos encontramos a menudo con los sistemas de ecuaciones que se llaman rígidos y cuya resolución por medios corrientes representa grandes dificultades. Supongamos que $\{\lambda_k\}$ son los números propios de la matriz A (si A no es simétrica, los números λ_k pueden ser complejos). El sistema de ecuaciones (33) se denominará *rígido*, si $\operatorname{Re}\lambda_k > 0$ $(k = 1, 2, \ldots, N)$ y si la razón $\xi = \max\limits_k \operatorname{Re}\lambda_k / \min\limits_k \operatorname{Re}\lambda_k$ es grande.

Si la matriz A es simétrica, entonces todos los números propios son reales y la rigidez del sistema (33) es testimonio de que la matriz A es positiva y que el sistema (33) está mal condicionado, es decir,

$$\xi = \frac{\max\limits_k \lambda_k}{\min\limits_k \lambda_k} \gg 1.$$

Son rígidas, en particular, las ecuaciones que se obtienen al reducir las ecuaciones con derivadas parciales a los siste-

mas de ecuaciones diferenciales ordinarias por medio de la aproximación de diferencias de un operador que contiene derivadas respecto de las variables espaciales (por ejemplo, el operador de Laplace en el caso de la ecuación de conductibilidad térmica).

Los métodos explícitos resultaron ser inútiles para la resolución numérica de los esquemas rígidos, puesto que conducen a grandes restricciones referentes al paso a causa de las exigencias de estabilidad en perjuicio de las de precisión. Aclaremos esto con un ejemplo del sistema de dos ecuaciones

$$\frac{du_1}{dt}+a_1u_1=0,\quad \frac{du_2}{dt}+a_2u_2=0,\quad A=\begin{pmatrix}a_1 & 0\\ 0 & a_2\end{pmatrix},\quad t>0$$

$$a_2>0,\quad a_1>0,\quad a_2\gg a_1. \tag{34}$$

La solución de este sistema es un vector

$$u(t)=(u_1(t),\ u_2(t)),$$

$$u_1(t)=u_1(0)\,e^{-a_1t},\ u_2(t)=u_2(0)\,e^{-a_2t};$$

los componentes de este vector decrecen cuando crece t, con la particularidad de que $|u_2(t)|\ll|u_1(t)|$ para t suficientemente grande.

Tomemos un esquema explícito

$$\frac{y_1^{n+1}-y_1^n}{\tau}+a_1y_1^n=0,\quad \frac{y_2^{n+1}-y_2^n}{\tau}+a_2y_2^n=0,$$

$$n=0,\ 1,\ \ldots,\ y_i^n=y_i(t_n),\quad i=1,\ 2. \tag{35}$$

El sistema se descompone en dos ecuaciones, cada una de las cuales puede resolverse separadamente, no obstante dichas ecuaciones están ligadas entre sí por la elección del paso común τ. El esquema es estable, si se cumplen simultáneamente dos condiciones $a_1\tau\leqslant 2$ y $a_2\tau\leqslant 2$. Por cuanto $a_2\gg a_1$, ambas condiciones quedan cumplidas, si $\tau\leqslant 2/a_2$. El paso admisible τ se determina, de hecho, por aquel componente $u_2(t)$ de la solución que decrece con mayor rapidez.

Para la resolución del sistema (34) es aplicable un esquema implícito

$$\frac{y_1^{n+1}-y_1^n}{\tau}+a_1y_1^{n+1}=0,\quad \frac{y_2^{n+1}-y_2^n}{\tau}+a_2y_2^{n+1}=0,$$

que es estable para cualesquiera τ y $a_1\geqslant 0$, $a_2\geqslant 0$.

Ultimamente ha aparecido toda una serie de esquemas implícitos, algoritmos para éstos y programas nuevos, útiles para resolver sistemas rígidos de ecuaciones diferenciales lineales y no lineales.

6. Observaciones generales. 1. Al escoger tal o cual método numérico se toman en consideración varias circunstancias, a saber, el volumen de los cálculos, el volumen requerido de memoria de acceso rápido del ordenador, el orden de exactitud, la estabilidad respecto de los errores de redondeo y otras. Hemos considerado en todo caso los métodos de paso constante $\tau = t_{n+1} - t_n$. La introducción del paso variable $\tau_{n+1} = t_{n+1} - t_n$ lleva un carácter formal y, cuando se trata de los esquemas de un paso, no conduce a nuevas cuestiones de principio. Para los esquemas de varios pasos ($m \geqslant 2$) las fórmulas se alteran.

En el caso general la solución puede ser una función no monótona fuertemente variable. Es natural de emplear una red no uniforme y disminuir el paso (espesar la red) en el dominio de variación rápida de la función $u(t)$ con el fin de asegurar una aproximación más exacta de $u(t)$ por la solución reticular. Sin embargo, no sabemos de antemano el comportamiento de la solución $u = u(t)$. Por eso en la práctica se procede de la manera siguiente: al principio se realizan los cálculos en la red uniforme; si se pone claro que la solución $u = u(t)$ varía fuertemente en cierto intervalo $t_* < t < t^*$, entonces la red se hace espesar en $[t_*, t^*]$ y el problema se resuelve en la red no uniforme de esta índole. Se recomienda, en general, realizar los cálculos en varias redes que se hacen espesar. Si, al espesar la red, la solución varía poco, la exactitud requerida se considera lograda. Con el fin de elevar el orden de exactitud resulta aplicable el método de Runge en el que se usan cálculos sobre diferentes redes (siempre que la solución $u = u(t)$ posee una suavidad suficiente). En el transcurso de los cálculos puede resultar necesario emplear los esquemas de diferentes órdenes de exactitud en distintos dominios de variación del argumento.

2. Nos encontramos a menudo con la necesidad de resolver las ecuaciones cuyos coeficientes cambian furtemente, por ejemplo

$$\frac{du}{dt} = \alpha(t)\, u, \quad t > 0, \quad u(0) = u_0. \qquad (36)$$

Una ecuación de este género se encuentra en la descripción de los problemas de la cinética química. A título de su solución interviene una función

$$u(t) = u_0 \exp\left\{\int_0^t \alpha(s)\, ds\right\}.$$

Si $\alpha(t) \geqslant 0$, puede utilizarse el esquema de Euler para cualquiera:

$$y_{n+1} = y_n + \tau\alpha_n y_n = (1 + \tau\alpha_n)\, y_n. \qquad (37)$$

Si, en cambio, $\alpha(t) < 0$, puede suceder que $1 + \tau\alpha_{n_1} < 0$ para cierto $n = n_1$, e $y_{n_1+1} < 0$, es decir, la solución pierde sentido. En este caso puede emplearse un esquema implícito

$$y_{n+1} = y_n + \tau\alpha_n y_{n+1},$$

$$y_{n+1} = y_n/(1 - \tau\alpha_n), \qquad 1 - \tau\alpha_n > 1, \qquad (38)$$

que es estable para cualquier τ. Si $\alpha(t)$ cambia de signo para ciertos valores de t, entonces en aquellos nodos, donde $\alpha(t) > 0$, se debe emplear el esquema explícito (37), y en los nodos en que $\alpha(t) < 0$, el esquema implícito (38).

Los métodos de Adams son menos laboriosos en comparación con los de Runge—Kutta. La deficiencia de los métodos de Adams radica en lo que el comienzo de los cálculos no es usual; para determinar $y_1, y_2, \ldots, y_{m-1}$ se emplea corrientemente el método de Runge—Kutta. Si se emplean los esquemas de Adams de dos pasos (y, con mayor razón, de varios pasos) el cambio del paso τ requiere cierta complicación de las fórmulas, lo que no ocurre al emplear el método de Runge—Kutta. En la práctica se emplea la combinación de los métodos de Runge—Kutta y de Adams con un programa de elección automática del paso para la obtención de la exactitud prefijada.

§ 3. Aproximación del problema de Cauchy para un sistema de ecuaciones diferenciales lineales ordinarias de primer orden

1. Problema de Cauchy. En este párrafo se estudiarán los esquemas de diferencias lineales (de un paso o de dos capas) que surgen al aproximar el problema de Cauchy para un

sistema de ecuaciones diferenciales lineales ordinarias de primer orden, como también al aproximar las ecuaciones diferenciales con derivadas parciales (método de las rectas).

Tomemos un problema de Cauchy

$$\frac{du^i}{dt}+\sum_{j=1}^{N} a_{ij}u^j=f^i(t),\quad t\geqslant 0,\quad u^i(0)=u_0^i,$$
$$i=1,\ 2,\ \ldots,\ N. \quad (1)$$

Al designar por $A=(a_{ij})$ una matriz cuadrada de dimensión $N\times N$ con elementos a_{ij} que no dependen de t, por $u(t)=(u^1(t),\ u^2(t),\ \ldots,\ u^N(t))$ y $f(t)=(f^1(t),\ f^2(t),\ \ldots,\ f^N(t))$ los vectores buscado y prefijado de dimensión N, respectivamente, escribamos el sistema en la forma

$$\frac{du}{dt}+Au=f(t),\quad t\geqslant 0,\quad u(0)=u_0. \quad (2)$$

La misma designación A se empleará también para un operador correspondiente que actúa en el espacio H^N de dimensión N (A: $H^N\to H^N$). Introduzcamos en el espacio H^N un producto escalar $(u,\ v)$ y una norma $\|u\|=\sqrt{(u,\ u)}$. Supondremos que el operador A es positivo

$$A>0,\ \text{ó}\ (Ax,\ x)>0\ \text{para todos los}\ x\in H^N,\ x\neq 0.$$

El problema de Cauchy (1) bajo las condiciones (2) tiene la solución única. En efecto, supongamos que existen dos soluciones $\bar{u}(t)$ y $\bar{u}(t)$ del problema (2). En este caso su diferencia satisface las condiciones homogéneas

$$\frac{dz}{dt}+Az=0,\quad t>0,\quad z(0)=0,\quad z(t)=\bar{u}(t)-\bar{\bar{u}}(t). \quad (3)$$

Multiplicando (3) escalarmente por z, y tomando en consideración que $\left(z,\ \frac{dz}{dt}\right)=\frac{1}{2}\frac{d}{dt}(z,\ z)$, obtenemos

$$\frac{1}{2}\frac{d}{dt}\|z\|^2+(Az,\ z)=0,$$

$$\|z(t)^2\| \quad \int_0^t (Az(t'),\ z(t'))\,dt'=\|z(0)\|^2.$$

Puesto que $A > 0$, $z(0) = 0$, de aquí se deduce que

$$\| z(t) \|^2 = 0, \quad z(t) \equiv 0, \quad \bar{u}(t) \equiv \bar{\bar{u}}(t).$$

Indiquemos una propiedad de importancia de la solución del problema (2) cuando $f(t) \equiv 0$:

$$\| u(t) \| \leqslant e^{-\lambda_1 t} \| u(0) \|, \text{ si } A = A^* > 0, \tag{4}$$

donde λ_1 es el valor propio mínimo del operador A:

$A\xi_k = \lambda_k \xi_k$, $k = 1, 2, \ldots, N$, $0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N$.

Con el fin de demostrar (4) buscaremos la solución $u(t)$ del problema (2) en la forma

$$u(t) = \sum_{k=1}^{N} \alpha_k(t)\, \xi_k, \qquad \|u(t)\|^2 = \sum_{k=1}^{N} \alpha_k^2(t).$$

Al sustituir esta expresión en la ecuación (2) con $f(t) \equiv 0$, hallamos

$$\sum_{k=1}^{N} \left(\frac{d\alpha_k}{dt} + \lambda_k \alpha_k \right) \xi_k = 0,$$

y, por lo tanto, $\frac{d\alpha_k}{dt} + \lambda_k \alpha_k = 0$, $\alpha_k(t) = \alpha_k(0)\, e^{-\alpha_k t}$, de suerte que

$$\|u(t)\|^2 = \sum_{k=1}^{N} \alpha_k^2(0)\, e^{-2\lambda_k t} \leqslant e^{-2\lambda_1 t} \sum_{k=1}^{N} \alpha_k^2(0) = \\ = e^{-2\lambda_1 t} \|u(0)\|^2.$$

2. Esquemas de diferencias. Introduzcamos una red de paso τ según la variable t: $\omega_\tau = \{t_n = n\tau, n = 0, 1, 2, \ldots\}$ y designemos con $y_n = y(t_n)$ una función reticular del argumento $t_n = n\tau$ (o bien n) con valores en H^N. Escribamos un esquema explícito

$$\frac{y_{n+1} - y_n}{\tau} + Ay_n = f_n, \quad n = 0, 1, 2, \ldots, \quad y_0 = u_0, \tag{5}$$

de modo que y_{n+1} se halla por la fórmula explícita

$$y_{n+1} = y_n - \tau(Ay_n - f_n),\ n = 0, 1, 2, \ldots,\ y_0 = u_0. \tag{5'}$$

La solución y_n del problema (5) depende no sólo de τ, sino también de N o del parámetro $h = 1/N$: $y_n = y_{n,\tau,h}$. En realidad analizamos no solo un problema (5), sino un *conjunto de problemas* $\{5_{\tau,h}\}$ *para* cualesquiera τ y h. Esto es precisamente un esquema de diferencias. A título de su solución interviene una familia de funciones $\{y_{n,\tau,h}\}$. Para no complicar las notaciones, omitiremos los índices τ y h en los casos cuando esto no lleve a equivocaciones algunas. El esquema (5) es *esquema de diferencias de un paso* (o *de dos capas*).

En general, por *esquema de dos capas* se entiende una ecuación que liga los valores del vector $y(t)$ para dos valores del argumento $t = t_n$ y $t = t_{n+1}$ (para dos capas):

$$By_{n+1} = Cy_n + F_n, \qquad n = 0, 1, \ldots,$$

donde B, C son las matrices cuadradas $N \times N$ (operadores lineales B, C: $H^N \to H^N$); y_n, F_n son los vectores de dimensión N. Dicha ecuación puede ser siempre escrita en la siguiente forma canónica:

$$B\frac{y_{n+1}-y_n}{\tau} + Ay_n = \varphi_n, \quad n=0, 1, 2, \ldots, \quad y_0 = u_0. \tag{6}$$

Para determinar y_{n+1} se debe resolver la ecuación

$$BY_{n+1} = \Phi_n, \quad \Phi_n = By_n - \tau(Ay_n - \varphi_n).$$

Supondremos siempre que existe el operador inverso B^{-1}.

Si $B = E$ es un operador unidad, obtendremos el *esquema explícito* (5). Cuando $B \neq E$, el esquema (6) se denomina *implícito*. Se encuentran con frecuencia los esquemas

$$\frac{y_{n+1}-y_n}{\tau} + Ay_{n+1} = \varphi_n \text{ (esquema explícito puro)}, \tag{7}$$

$$\frac{y_{n+1}-y_n}{\tau} + \frac{1}{2}A(y_n + y_{n+1}) = \varphi_n \text{ (esquema simétrico)}. \tag{8}$$

Representan ambos los casos particulares (para $\sigma = 1$ y $\sigma = 1/2$) del *esquema con pesos*

$$\frac{y_{n+1}-y_n}{\tau} + A(\sigma y_{n+1} + (1-\sigma)y_n) = \varphi_n, \quad u = 0, 1, \ldots, \tag{9}$$

el cual puede escribirse en la forma canónica (6) con

$$B = E + \sigma\tau A, \tag{10}$$

si tenemos en cuenta que $\sigma y_{n+1} + (1-\sigma) y_n = y_n + \sigma\tau (y_{n+1} - y_n)/\tau$.

3. Error de aproximación. Sea $u = u(t)$ la solución del problema (2) y sea $y_n = y(t_n)$ la solución del problema (6); al sustituir en (6) $y_n = u_n + z_n$, obtenemos para el error $z_n = y_n - u_n$, $u_n = u(t_n)$:

$$B\frac{z_{n+1}-z_n}{\tau} + Az_n = \psi_n, \quad n = 0, 1, 2, \ldots, z_0 = 0, \quad (11)$$

donde

$$\psi_n = \varphi_n - Au_n - B\frac{u_{n+1}-u_n}{\tau} \quad (12)$$

es el residuo o error de aproximación para el esquema (6) en la solución $u = u(t)$ del problema de partida (2).

Sean $\| u \|_1$, $\| v \|_{(2)}$ ciertas normas en $H^N = H_h$. El esquema (6) converge, si $\| z_n \|_{(1)} \to 0$ para $\tau \to 0$, cualquiera que sea $n = 1, 2, \ldots$. El esquema (6) es de *m-ésimo orden de exactitud*, o converge con la velocidad $O(\tau^m)$, siempre que

$$\| z_n \|_{(1)} = O(\tau^m), \text{ es decir } \| z_n \|_{(1)} \leqslant M\tau^m, \quad (13)$$

donde $M = \text{const}$ no depende de τ.

Recordemos que el esquema (6) es de *m-ésimo orden de aproximación* en la solución de la ecuación (1), si para el residuo ψ_n se cumple la estimación

$$\| \psi_n \|_{(2)} = O(\tau^m). \quad (14)$$

Aclaremos las condiciones de aproximación del esquema (6) con $m = 1, 2$. Suponiendo que $u = u(t)$ tiene tantas derivadas cuanto sea necesario en el transcurso de exposición, encontramos

$$u_{n+1} = \left(u + \frac{\tau}{2}\dot{u} + \frac{\tau^2}{8}\ddot{u}\right)_{n+1/2} + O(\tau^3),$$

$$\dot{u}_n = \left(\frac{du}{dt}\right)_n, \quad \ddot{u}_n = \left(\frac{d^2u}{dt^2}\right)_n,$$

$$u_n = \left(u - \frac{\tau}{2}\dot{u} + \frac{\tau^2}{8}\ddot{u}\right)_{n+1/2} + O(\tau^2),$$

$$\frac{1}{\tau}(u_{n+1} - u_n) = \dot{u}_{n+1/2} + O(\tau^2),$$

$$\psi_n = \varphi_n - (Au + B\dot{u})_{n+1/2} + \frac{\tau}{2} A\dot{u}_{n+1/2} + O(\tau^2) =$$
$$= \varphi_n - f_{n+1/2} + (f - Au - \dot{u})_{n+1/2} +$$
$$+ \left(E - B + \frac{\tau}{2} A\right) \dot{u}_{n+1/2} + O(\tau^2) =$$
$$= \varphi_n - f_{n+1/2} + \left(E - B + \frac{\tau}{2} A\right) \dot{u}_{n+1/2} + O(\tau^2).$$

De aquí se ve que la condición (14) se cumplirá, si

$$\|\varphi_n - f_{n+1/2}\|_{(2)} = O(\tau^m),$$
$$\left\| \left(E - B + \frac{\tau}{2} A\right) \dot{u} \right\|_{(2)} = O(\tau^m), \quad m = 1,\ 2. \tag{15}$$

En particular, para el esquema explícito (en el caso $B = E$) tenemos

$$\left\| \frac{\tau}{2} A\dot{u} \right\|_{(2)} = O(\tau),$$

y $\| \psi_n \|_{(2)} = O(\tau)$ cuando $\| \varphi_n - f_{n+1/2} \| = O(\tau)$, por ejemplo, cuando $\varphi_n = f_n$.

Si, en el caso del esquema simétrico ($\sigma = 1/2$) $B = E + \tau A/2$, $\| \varphi_n - f_{n+1/2} \|_{(2)} = O(\tau^2)$, entonces $\| \psi_n \|_{(2)} = O(\tau^2)$, puesto que $\| (E - B + \tau A/2)\, \dot{u} \|_{(2)} = 0$; en tal caso podemos tomar, por ejemplo, $\varphi_n = f_{n+1/2}$.

El esquema de adelantamiento ($\sigma = 1$) es de primer orden de aproximación, puesto que $\|(E - B + \tau A/2)\, \dot{u} \|_{(2)} = \tau \| A\dot{u} \|_{(2)}/2 = O(\tau)$.

4. Estabilidad y convergencia. Según se ha observado más arriba, el esquema (6) es *estable* (respecto de los datos iniciales y del segundo miembro), si su solución depende continuamente de los datos de entrada (de y_0 y de φ_n), con la particularidad de que dicha dependencia es continua según τ y N, o según h. Para estimar la solución del problema se usará la norma $\| u \|_{(1)}$, y para estimar el segundo miembro, la norma $\| v \|_2$. Hagamos uso de la definición de estabilidad.

El esquema (6) será *estable*, si para cualesquiera y_0, φ_n existen tales constantes $M_1 > 0$ y $M_2 > 0$, independientes

tanto de τ, como de N, y_0, φ_n, que para la solución del problema (6) se cumple la desigualdad

$$\| y_n \|_{(1)} \leqslant M_1 \| y_0 \|_{(1)} + M_2 \max_{0 \leqslant k < n} \| \varphi_k \|_{(2)}. \quad (16)$$

Si el esquema (6) es estable y posee una aproximación $\| \psi_n \|_{(2)} \to 0$ cuando $\tau \to 0$, es convergente:

$$\| y_n - u_n \|_{(1)} \to 0 \text{ cuando } \tau \to 0, \; n = 1, 2, \ldots \quad (17)$$

(de la aproximación y de la estabilidad se desprende la convergencia del esquema). En efecto, si el esquema (6) es estable, entonces para la solución $z_n = y_n - u_n$ del problema (11) se cumple, de acuerdo con (16), la estimación

$$\|z_n\|_{(1)} \leqslant M_1 \max_{0 \leqslant k < n} \|\psi_k\|_{(2)}. \quad (18)$$

De aquí precisamente proviene que $\| z_n \|_{(1)} \to 0$, si $\| \psi_n \|_{(2)} \to$ $\to 0$ cuando $\tau \to 0$.

El estudio de la convergencia y del orden de exactitud se reduce al estudio del error de aproximación y de estabilidad del esquema de diferencias (6).

§ 4. Estabilidad del esquema de dos capas

1. Estabilidad respecto de los datos iniciales. Examinaremos un esquema de dos capas en la forma canónica

$$B \frac{y_{n+1} - y_n}{\tau} + A y_n = \varphi_n, \quad n = 0, 1, \ldots,$$

$$\text{se ha prefijado el valor inicial } y_0 \in H, \quad (1)$$

donde A, B: $H \to H$ $(H = H^N)$. La solución del problema (1) puede ser representada como una suma $y = y^{(1)} +$ $+ y^{(2)}$ de las soluciones de dos problemas

$$B \frac{y_{n+1} - y_n}{\tau} + A y_n = 0, \qquad n = 0, 1, \ldots, y_0 = u_0, \quad (2)$$

$$B \frac{y_{n+1} - y_n}{\tau} + A y_n = \varphi_n, \quad n = 0, 1, \ldots, y_0 = 0, \quad (3)$$

($y^{(1)}$ es la solución del problema (2), $y^{(2)}$ es la solución del problema (3)).

El esquema (1) es estable respecto de *los datos iniciales*, si para la solución del problema (2) es justa la estimación

$$\| y_n \|_{(1)} \leqslant M_1 \| y_0 \|_{(1)}. \qquad (4)$$

El esquema (1) es estable respecto del segundo miembro, si para la solución del problema (3) es justa la estimación

$$\|y_n\|_{(1)} \leqslant M_2 \max_{0 \leqslant k < n} \|\varphi_k\|_{(2)}. \qquad (5)$$

Aquí M_1, M_2 no dependen de N, τ, n.

Utilizaremos una condición más simple para la estabilidad respecto de los datos iniciales:

$$\| y_n \|_{(1)} \leqslant \| y_{n-1} \|_{(1)}, \ldots, \| y_1 \|_{(1)} \leqslant \leqslant \| y_0 \|_{(1)} \ (M_1 = 1), \qquad (6)$$

y, además, la condición de ρ-estabilidad:

$$\| y_n \|_{(1)} \leqslant \rho \| y_{n-1} \|_{(1)} \leqslant \ldots \leqslant \rho^n \| y_0 \|_{(1)}, \quad \rho > 0. \qquad (7)$$

Es evidente que el esquema es estable en el sentido de la definición (4), si $\rho = e^{c_0 \tau}$, donde $c_0 =$ const no depende de n, τ, N. En este caso $\rho^n = e^{c_0 t_n} \leqslant e^{c_0 T} = M_1$ para $0 \leqslant t_n \leqslant T$, $c_0 > 0$, o bien $\rho^n \leqslant 1$ para $c_0 \leqslant 0$.

Introduzcamos en el espacio H un producto escalar (,) y una norma $\| x \| = \sqrt{(x, x)}$. Sea $D = D^* > 0$ un operador positivo *autoconjugado*. A título de norma $\| y \|_{(1)}$ elijamos una norma energética

$$\|y\|_{(1)} = \|y\|_D = \sqrt{(Dy, y)}. \qquad (8)$$

En particular, $D = A$, $D = E$ o $D = B$ (para $B = B^* > 0$). De (2) proviene que

$$y_{n+1} = Sy_n, \quad S = E - \tau B^{-1}A, \qquad (9)$$

donde S es el operador de transición de una capa a la otra.

El esquema (2) es estable en H_D, si es justa la estimación

$$\| y_{n+1} \|_D \leqslant \| y_n \|_D. \qquad (10)$$

De la estimación $\| y_{n+1} \|_D = \| Sy_n \|_D \leqslant \| S \|_D \| y_n \|_D$ se deduce que la desigualdad (10) es equivalente a la condición

$$\| S \|_D \leqslant 1. \qquad (11)$$

Esta última condición es equivalente, a su vez, a la siguiente

$$J_D = \| y \|_D^2 - \| Sy \|_D^2 = (Dy, y) - (DSy, Sy) \geqslant 0$$
para todo $y \in H$. (12)

De este modo, (10), (11) y (12) son equivalentes, es decir, el cumplimiento de cualquiera de ellas provoca el cumplimiento de dos otras.

2. Condición necesaria y suficiente de estabilidad. Teorema fundamental.

TEOREMA 1. *Si $A = A^*$ es un operador positivo autoconjugado y existe el operador B^{-1}, entonces para que el esquema (2) sea estable en H_A:*

$$\| y_{n+1} \|_A \leqslant \| y_n \|_A \quad (13)$$

es necesario y suficiente que se verifique la desigualdad

$$(By, y) - \frac{\tau}{2} (Ay, y) \geqslant 0 \text{ para todo } y \in H, \text{ o bien } B \geqslant \frac{\tau}{2} A. \quad (14)$$

DEMOSTRACION. Es suficiente convencerse de equivalencia entre (14) y la desigualdad $J_A \geqslant 0$, donde

$$J_A = (Ay, y) - (ASy, Sy) = \\ = (Ay, y) - (Ay - \tau AB^{-1}Ay, y - \tau B^{-1}Ay) = \\ = 2\tau (AB^{-1}Ay, y) - \tau^2 (AB^{-1}Ay, B^{-1}Ay).$$

Al designar $B^{-1}Ay = x$, $Ay = Bx$, obtendremos

$$J_A = 2\tau \left((Bx, x) - \frac{\tau}{2} (Ax, x) \right) \geqslant 0 \text{ para todo } x \in H, \quad (15)$$

es decir, las desigualdades (14), (15) y, por lo tanto, (13), (14) son equivalentes. Esto quiere decir que de (14) provienen (11), (12) para $D = A$ y (13) (la condición (14) es suficiente para la estabilidad). Por cuanto el esquema es estable, es decir, se verifica (13) o bien $\| S \|_A \leqslant 1$, entonces $J_A \geqslant 0$, y, por consiguiente, $B \geqslant \tau A/2$ (necesidad de la condición (14)).

OBSERVACION. La condición (14) puede aclararse con un ejemplo del esquema de diferencias

$$b \frac{y_{n+1} - y_n}{\tau} + a y_n = 0, \; n = 0, 1, 2, \ldots, \; a > 0, \; b > 0$$

con coeficientes numéricos a, b. Este esquema corresponde al problema de Cauchy

$$bu'(t) + au(t) = 0 \quad t > 0, \ u(0) = u_0.$$

De la fórmula $y_{n+1} = (1 - \tau a/b)\, y_n$ se ve que el esquema es estable, es decir, $|y_{n+1}| \leqslant |y_n| \leqslant \ldots \leqslant |y_0|$, si $|1 - \tau a/b| \leqslant 1$, $-1 \leqslant 1 - \tau p/b \leqslant 1$, es decir, $b \geqslant \tau a/2$. La analogía con la desigualdad operacional $B \geqslant \tau A/2$ es evidente.

3. Ejemplos de aplicación del teorema fundamental.
EJEMPLO 1. Un esquema explícito: $B = E$, $A = A^* > 0$. De la desigualdad de Cauchy—Buniakovski $(Ax, x) \leqslant \leqslant \|Ax\| \|x\| \leqslant \|A\| \|x\|^2$ se deduce $A \leqslant \|A\| E$, o bien

$$E \geqslant \frac{1}{A} A \tag{16}$$

Veamos ahora la diferencia $B - \frac{1}{2}\tau A = E - \frac{1}{2}\tau A \geqslant \frac{1}{\|A\|} A - \frac{1}{2}\tau A = \left(\frac{1}{\|A\|} - \frac{\tau}{2}\right) A$. Como que $A > 0$, entonces la condición $B - \frac{1}{2}\tau A \geqslant 0$ se cumplirá para $\frac{1}{\|A\|} - \frac{\tau}{2} \geqslant 0$, es decir, cuando

$$\tau \leqslant 2/\|A\|. \tag{17}$$

Esta es una condición necesaria y suficiente de estabilidad del esquema explícito en H_A ($\|y_n\|_A \leqslant \|y_0\|_A$).

EJEMPLO 2. Esquema (9) del § 3 con pesos, $A = A^* > 0$. Para dicho esquema $B = E + \sigma\tau A$ y $B - \frac{1}{2}\tau A = E + \left(\sigma - \frac{1}{2}\right)\tau A \geqslant \frac{1}{\|A\|} + \left(\left(\sigma - \frac{1}{2}\right)\tau\right) A \geqslant 0$, siempre que

$$1 + \left(\sigma - \frac{1}{2}\right)\tau\|A\| \geqslant 0. \tag{18}$$

De aquí se ve que el esquema con pesos es estable en H_A para todo $\tau > 0$ (incondicionalmente estable), si $\sigma \geqslant 1/2$, y es condicionalmente estable para $\tau \leqslant 1/(1/2 - \sigma)\|A\|$, si $\sigma < 1/2$.

EJEMPLO 3. Estabilidad en H (para $D = E$) del esquema con pesos (9) del § 3:

$$(E + \sigma\tau A)\frac{y_{n+1} - y_n}{\tau} + Ay_n = 0, \quad n = 0, 1, 2, \ldots,$$
$$B = E + \sigma\tau A. \tag{19}$$

Aplicando el operador A^{-1} a los dos miembros de la ecuación (19), obtenemos

$$\widetilde{B}\frac{y_{n+1} - y_n}{\tau} + \widetilde{A}y_n = 0, \quad n = 0, 1, 2, \ldots,$$
$$\widetilde{B} = A^{-1} + \sigma\tau E, \ \widetilde{A} = E, \tag{20}$$

Este esquema es estable, en virtud del teorema 1, en $H_{\widetilde{A}} = H$ $(\widetilde{A}^* = \widetilde{A} = E > 0)$ para $\widetilde{B} - \frac{1}{2}\tau\widetilde{A} = A^{-1} + \left(\sigma - \frac{1}{2}\right)\tau E \geqslant \left(\frac{1}{\|A\|} + \left(\sigma - \frac{1}{2}\right)\tau\right) E \geqslant 0$, es decir, si se cumple (18) (en este caso se ha tomado en consideración la estimación $A^{-1} \geqslant \frac{1}{\|A\|}E$, la cual se deduce de (16)). De este modo, de (18) se infiere que para (19) es justa la estimación (10) cuando $D = \widetilde{A}$, es decir,

$$\|y_n\| \leqslant \|y_0\|. \tag{21}$$

El esquema (19) puede ser escrito en la forma

$$y_{n+1} = Sy_n, \ S = (E + \sigma\tau A)^{-1}) (E - (1 - \sigma)\tau A),$$
$$A = A^* > 0. \tag{22}$$

Por eso, si se cumple la condición (18), para dicho esquema es justa la estimación (21), lo que significa

$$\|(E + \sigma\tau A)^{-1}(E - (1 - \sigma)\tau A)\| \leqslant 1,$$
$$\text{siempre que } 1 + \left(\sigma - \frac{1}{2}\right)\tau\|A\| \geqslant 0. \tag{23}$$

Esta estimación nos hará falta en lo que sigue.

4. Estabilidad en H .

TEOREMA 2. *Si* $A = A^* > 0$, $B = B^* > 0$, *entonces para que el esquema* (2) *sea estable en* H_B:

$$\|y_{n+1}\|_B \leqslant \|y_n\|_B, \tag{24}$$

es necesario y suficiente que se cumpla la condición (14).

DEMOSTRACIÓN. Escribamos el esquema (2) en la forma (9) y mostremos que la condición

$$\| S \|_B \leqslant 1 \tag{25}$$

es equivalente a la desigualdad (14), es decir, de (14) proviene (25), y, viceversa, de (25) proviene (14).

Sea y un vector arbitrario de H; representémoslo en la forma

$$y = \sum_{k=1}^{N} \alpha_k \xi_k,$$

donde $\{\xi_k\}$ son los vectores propios del problema

$$A\xi_k = \lambda_k B\xi_k, \qquad \lambda_k > 0,$$
$$(B\xi_k, \xi_m) = \delta_{km} = \begin{cases} 1, & k = m, \ k, m = 1, 2, \ldots, N \\ 0, & k \neq m. \end{cases} \tag{26}$$

Teniendo presente que $S\xi_k = \xi_k - \tau B^{-1}A\xi_k = (1 - \tau\lambda_k)\xi_k$, $BS\xi_k = (1 - \tau\lambda_k)B\xi_k$, encontramos

$$(By, y) = \sum_{k=1}^{N} \alpha_k^2, \qquad (Ay, y) = \sum_{k=1}^{N} \lambda_k \alpha_k^2,$$
$$(BSy, \ Sy) = \sum_{k=1}^{N} \alpha_k^2 (1 - \tau\lambda_k)^2 \leqslant \|S\|_B^2 \sum_{k=1}^{N} \alpha_k^2 = \|S\|_B^2 (By, y), \tag{27}$$

donde

$$\|S\|_B^2 = \max_{1 \leqslant k \leqslant N} (1 - \tau\lambda_x)^2, \tag{28}$$

La desigualdad (25) es equivalente a la condición

$$\tau\lambda_k \leqslant 2, \qquad k = 1, 2, \ldots, N, \tag{29}$$

la cual, a su vez, es equivalente a la desigualdad (14), puesto que

$$(By, y) - \frac{\tau}{2}(Ay, y) = \sum_{k=1}^{N} \alpha_k^2 \left(1 - \frac{\tau\lambda_k}{2}\right).$$

Con ello queda demostrada la equivalencia de (24) y (14).

5. ρ-estabilidad.

TEOREMA 3. *Si* $A = A^* > 0$, $B = B^* > 0$, *entonces la condición necesaria y suficiente de la* ρ*-estabilidad del esquema* (2) *con* $\rho > 0$ *cualquiera:*

$$\| y_{n+1} \|_D \leqslant \rho \| y_n \|_D, \quad D = A, B, \tag{30}$$

será representada por las desigualdades operacionales

$$\frac{1-\rho}{\tau} B \leqslant A \leqslant \frac{1+\rho}{\tau} B. \tag{31}$$

DEMOSTRACIÓN. Las desigualdades (31) son equivalentes a las condiciones (véase el cap. I, § 4, p. 4):

$$\frac{1-\rho}{\tau} \leqslant \lambda_k \leqslant \frac{1+\rho}{\tau}, \quad k=1, 2, \ldots, N, \tag{32}$$

donde λ_k son los números propios del problema (26).

Admitamos que $D = B$ y son justas (31) ó (32). De (32) se deduce $-\rho \leqslant \tau\lambda_k - 1 \leqslant \rho$, $|1 - \tau\lambda_k| \leqslant \rho$, y, en virtud de (27), $\| S \|_B \leqslant \rho$ (puesto que $\| S \|_B^2$ es una constante mínima, para la cual se verifica la desigualdad $(BSy, Sy) \leqslant \leqslant M (By, y)$, es decir, es justa la estimación (30) suficiencia). Si es justa la estimación (30), entonces $|1 - \tau\lambda_k| \leqslant \leqslant \rho$, y, por lo tanto, quedan cumplidas (32) y (31) (necesidad).

Análogamente se demuestra el teorema para $D = A$, si se toma en consideración que

$$(ASy, Sy) = \sum_{k=1}^{N} \alpha_k^2 \lambda_k (1-\tau\lambda_k)^2 \leqslant \max_{1 \leqslant k \leqslant N} (1-\tau\lambda_k)^2 (Ay, y).$$

De (30) se infiere

$$\| y_n \|_D \leqslant \rho^n \| y_0 \|_D.$$

Surge una pregunta: ¿en qué condiciones tiene lugar la estimación apriorística (30) con $\rho < 1$? La respuesta se ofrece por el siguiente

TEOREMA 4. Supongamos cumplidas las condiciones

$$A = A^* > 0, \; B = B^* > 0, \; \gamma_1 B \leqslant A \leqslant \gamma_2 B, \; \gamma_1 > 0. \tag{33}$$

En este caso para la solución del problema (2) *es justa la estimación*

$$\| y_{n+1} \|_D \leqslant \rho \| y_n \|_D, \quad \rho = 1 - \tau\gamma_1, \quad D = A, B, \tag{34}$$

siempre que

$$\tau \leqslant \tau_0, \qquad \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \tag{35}$$

Para la demostración se debe calcular la norma $\| S \|_B = \| S \|_A = \max_{1 \leqslant k \leqslant N} | 1 - \tau\lambda_k |$ **bajo la condición de que** $\gamma_1 \leqslant \leqslant \lambda_k \leqslant_2$, $0 < \gamma_1 = \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_N = \lambda_2$. **Veamos una diferencia** $\varphi_k = (1 - \tau\lambda_1)^2 - (1 - \tau\lambda_k)^2 = 2\tau (\lambda_k - \lambda_1) \times \times \left(1 - \frac{\tau}{2} (\lambda_k - \lambda_1)\right)$. **De aquí se ve que** $\varphi_k \geqslant 0$ **para** $1 - -\frac{\tau}{2} (\lambda_k + \lambda_1) \geqslant 1 - \frac{\tau}{2} (\gamma_2 + \gamma_1) \geqslant 1 - \frac{\tau_0}{2} (\gamma_1 - \gamma_2) = 0$, **es decir,** $\max_{1 \leqslant k \leqslant N} | 1 - \tau\lambda_k | = 1 - \tau\gamma_1$, **si** $\tau \leqslant \tau_0$. **El teorema está demostrado.**

6. Estabilidad respecto del segundo miembro. Método de las desigualdades energéticas. Analicemos el problema y reescribámoslo en la forma

$$y_{n+1} + Sy_n + \tau B^{-1}\varphi_n, \quad n = 0, 1, \ldots,$$
$$S = E - \tau B^{-1}A, \quad y_0 = 0. \tag{36}$$

Hagamos uso de la desigualdad triangular

$$\| y_{n+1} \|_D \leqslant \| Sy_n \|_D + \tau \| B^{-1}{}_n \|_D \leqslant$$
$$\leqslant \| S \|_D \| y_n \|_D + \tau \| B^{-1}\varphi_n \|_D. \tag{37}$$

Si se cumplen las condiciones del teorema 2, entonces $B = B^* > 0$, $D = B$, y $\| S \|_D = \| S \|_B \leqslant 1$ para $B \geqslant \frac{\tau}{2} A$, $\| B^{-1}\varphi_n \|_B^2 = (B (B^{-1}\varphi_n), B^{-1}\varphi_n) = (B^{-1}\varphi_n, \varphi_n) = \| \varphi_n \|_{B^{-1}}^2$, y de (37) proviene

$$\| y_{n+1} \|_B \leqslant \| y_n \|_B + \tau \| \varphi_n \|_{B^{-1}}.$$

Al sumar según $n = 0, 1, 2 \ldots$ y teniendo presente que $y_0 = 0$, obtendremos

$$\| y_n \|_B \leqslant \sum_{k=0}^{n-1} \tau \| \varphi_k \|_{B^{-1}}. \tag{38}$$

Esta estimación apriorística expresa la estabilidad del esquema (1) respecto del segundo miembro bajo la misma condición (14).

Pueden obtenerse también otras estimaciones. Con este fin aprovechemos un método, bastante general, de desigualdades energéticas. Sustituyamos $y_n = \frac{1}{2}(y_n + y_{n+1}) - \frac{\tau}{2} \times \times \frac{y_{n+1} - y_n}{\tau}$ en (1):

$$\left(B - \frac{\tau}{2} A\right) \frac{y_{n+1} - y_n}{\tau} + \frac{1}{2} A (y_{n+1} + y_n) = \varphi_n.$$

Multipliquemos esta ecuación escalarmente por $2(y_{n+1} - y_n)$ y tengamos en cuenta que $(A(y_{n+1} + y_n),\ y_{n+1} - y_n) = (Ay_{n+1}, y_{n+1}) + (Ay_n, y_{n+1}) + (Ay_{n+1}, y_n) - (Ay_n, y_n) = (Ay_{n+1},\ y_{n+1}) - (Ay_n,\ y_n)$, puesto que $(Ay_n,\ y_{n+1}) = (Ay_{n+1}, y_n)$ en virtud de que A es autoconjugado. De resultas se obtiene una «identidad energética»

$$2\tau\left(\left(B - \frac{\tau}{2} A\right) \frac{y_{n+1} - y_n}{\tau},\ \frac{y_{n+1} - y_n}{\tau}\right) + (Ay_{n+1},\ y_{n+1}) = \\ = (Ay_n,\ y_n) + 2(\varphi_n,\ y_{n+1} - y_n). \quad (39)$$

De aquí se ve que para $\varphi_n = 0$ y $B \geqslant \frac{\tau}{2} A$ será justa la estimación (13).

Transformemos $2(\varphi_n,\ y_{n+1} - y_n) = 2\tau\left(\varphi_n,\ \frac{y_{n+1} - y_n}{\tau}\right)$. Con este fin hagamos uso de la desigualdad

$$|ab| = (\sqrt{2\varepsilon}a)\left(\sqrt{\frac{1}{2\varepsilon}}b\right) \leqslant \varepsilon a^2 + \frac{1}{4\varepsilon} b^2,$$

donde a, b, $\varepsilon > 0$ son unos números cualesquiera. En nuestro caso

$$2(\varphi_n,\ y_{n+1} - y_n) \leqslant 2\tau \|\varphi_n\| \left\|\frac{y_{n+1} - y_n}{\tau}\right\| \leqslant \\ \leqslant 2\tau\varepsilon \left\|\frac{y_{n+1} - y_n}{\tau}\right\|^2 + \frac{\tau}{2\varepsilon}\|\varphi_n\|^2.$$

Al sustituir esta estimación en la identidad (39), obtendremos

$$2\tau\left(\left(B - \varepsilon E - \frac{\tau}{2} A\right) \frac{y_{n+1} - y_n}{\tau},\ \frac{y_{n+1} - y_n}{\tau}\right) + \\ + \|y_{n+1}\|_A^2 \leqslant \|y_n\|_A^2 + \frac{\tau}{2\varepsilon}\|\varphi_n\|^2. \quad (40)$$

Si se cumple la desigualdad

$$B \geqslant \varepsilon E + \frac{\tau}{2} A, \qquad \varepsilon > 0, \qquad (41)$$

entonces de (40) proviene (sustituyendo n por k)

$$\| y_{k+1} \|_A^2 \leqslant \| y_k \|_A^2 + \frac{\tau}{2\varepsilon} \| \varphi_k \|^2.$$

Al sumar según $k = 0, 1, 2, \ldots, n-1$, obtenemos una estimación

$$\| y_n \|_A^2 \leqslant \| y_0 \|_A^2 + \frac{1}{2\varepsilon} \sum_{k=0}^{n-1} \tau \| \varphi_k \|^2, \qquad (42)$$

que expresa la estabilidad del esquema (1) respecto del segundo miembro y de los datos iniciales en H_A.

EJEMPLO. Esquema con pesos (1): $B = E + \sigma\tau A$. Para dicho esquema la condición (41) significa que

$$(1 - \varepsilon) E + \left(\sigma - \frac{1}{2}\right) \tau A \geqslant 0.$$

En particular, la estimación (42) es justa para $\varepsilon = 1$ y $\sigma \geqslant 1/2$.

7. **Estabilidad asintótica.** Para el problema de Cauchy

$$\frac{du}{dt} + Au = 0, \qquad t > 0 \qquad u(0) = u_0$$

se ha obtenido en el § 3, p. 1, la estimación

$$\| u(t) \| \leqslant e^{-\gamma_1 t} \| u(0) \|,$$

donde $\lambda_1 = \min\limits_k \lambda_k (A)$.

Buscaremos las condiciones, bajo las cuales la estimación análoga tiene lugar para el esquema (2). Usaremos el teorema 4. Supongamos cumplidas las condiciones (33). Entonces, debido a (34), (35)

$$\| y_n \| A \leqslant \rho^n \| y_0 \|_A, \quad \rho = 1 - \tau\gamma_1, \quad \tau \leqslant \tau_0 = \frac{2}{\gamma_1 + \gamma_2}. \qquad (43)$$

De aquí se desprende una estimación que expresa la propiedad de la estabilidad asintótica

$$\| y_n \|_A \leqslant e^{-\gamma_1 t_n} \| y_0 \|_A \qquad (44)$$

(aquí se ha tomado en consideración que $\rho = 1 - \tau\gamma_1 < e^{-\tau\gamma_1}$).

Analicemos el esquema con pesos y supongamos que

$$\delta E \leqslant A \leqslant \Delta E, \quad \delta = \lambda_1 > 0, \quad \Delta = \lambda_N > 0. \quad (45)$$

Calculemos γ_1 y γ_2. Teniendo presente (45), tenemos

$$B = E + \sigma\tau A \geqslant \left(\frac{1}{\Delta} + \sigma\tau\right) A = \frac{1}{\gamma_2} A;$$

$$B \leqslant \left(\frac{1}{\delta} + \sigma\tau\right) A = \frac{1}{\gamma_1} A, \quad (46)$$

$$\gamma_1 = \frac{\delta}{1+\sigma\tau\delta}, \qquad \gamma_2 = \frac{\Delta}{1+\sigma\tau\Delta}.$$

Para un esquema explícito $\gamma_1 = \delta$, $\gamma_2 = \Delta$ la condición de estabilidad asintótica

$$\tau \leqslant 2/(\delta + \Delta) \quad (47)$$

es próxima a la condición de estabilidad usual con $\rho = 1$. Cuando $\sigma \neq 0$, la condición $\tau \leqslant 2/(\gamma_1 + \gamma_2)$ conduce a una desigualdad

$$2 + 2\,(\sigma - 1/2)\,\tau\,(\delta + \Delta) - 2\,\sigma\,(1 - \sigma)\,\tau^2\delta\Delta \geqslant 0.$$

Cuando $\sigma = 1$, se cumple para cualquier τ, es decir, un esquema implícito puro con $\sigma = 1$ es incondicionalmente estable asintóticamente. El esquema simétrica

$$\frac{y_{n+1} - y_n}{\tau} + \frac{1}{2} A\,(y_{n+1} + y_n) = 0, \quad \sigma = \frac{1}{2}, \quad (48)$$

es asintóticamente estable a condición de que

$$\tau \leqslant \tau^*, \quad \tau^* = 2/\sqrt{\delta\Delta}, \quad (49)$$

y incondicionalmente estable en el sentido habitual. En este caso

$$\rho = e^{-\lambda_1\tau + O(\tau^3)} < -e^{\lambda_1\tau}$$

y resulta lícita la estimación

$$\| y_n \| \leqslant e^{-\lambda_1 t_n} \| y_0 \| \quad \text{para} \quad \tau \leqslant \tau^*, \quad \sigma = 1/2. \quad (50)$$

¿Qué sucederá, si la condición $\tau \leqslant \tau_0$ no se cumple, es decir, si $\tau > \tau_0$? Entonces, máx $| 1 - \tau\lambda_k |$ se logra no para $k = 1$, sino para $k = N$ y $\rho = \tau\gamma_2 - 1$. La asintótica de la solución de un problema de diferencias (con t_n grandes) nada tiene en común con la solución asintótica del problema de partida. De este modo, la perturbación de la estabilidad asintótica lleva a la pérdida de la exactitud del esquema para t grandes.

Capítulo VI

Métodos de diferencias para las ecuaciones elípticas

En este capítulo se examinarán los esquemas de diferencias y los métodos de resolución de las ecuaciones en diferencias para la ecuación de Poisson y ecuaciones elípticas de coeficientes variables.

§ 1. Esquemas de diferencias para la ecuación de Poisson

1. Problema de partida. Examinemos la ecuación de Poisson

$$\Delta u = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2} = -f(x_1, x_2). \qquad (1)$$

Buscaremos su solución que sea continua en un rectángulo

$$\overline{G} = G \cup \Gamma = \{x = (x_1, x_2):\ 0 \leqslant x_\alpha \leqslant l_\alpha, \quad \alpha = 1, 2\}$$

y que tome en la frontera Γ los valores prefijados:

$$u\,|\Gamma = \mu(x) \qquad (2)$$

Un *problema* definido por la ecuación (1) y la condición (2). recibe el nombre *de Dirichlet* (*primer problema de contorno*).

2. Esquema de diferencias «cruz». Para la resolución numérica del problema (1), (2), introduzcamos en $\overline{G}$ una red $\overline{\omega}_h = \omega_h \cup \gamma_h = \{x_i = (i_1h_1, i_2h_2),\ i_\alpha = 0, 1, \ldots, N\alpha,$ $h_\alpha = l_\alpha/N_\alpha,\ \alpha = 1, 2\}$ y designemos con $y_i = y_{i_1 i_2} = y(i_1, i_2) = y(x_i)$ una función reticular definida sobre $\overline{\omega}_h$; h_1 y h_2 son los pasos de la red según las coordenadas de x_1 y x_2.

Con el fin de escribir un esquema de diferencias para (1), (2) aproximemos cada una de las derivadas $\partial^2 u/\partial x_\alpha^2$ en un

molde tripuntual, suponiendo

$$\frac{\partial^2 u}{\partial x_1^2} \sim \frac{u(x_1-h_1,\ x_2)-2u(x_1+x_2)+u(x_1+h_1,\ x_2)}{h_1^2}=u_{\bar{x}_1x_1},$$

$$\frac{\partial^2 u}{\partial x_2^2} \sim \frac{u(x_1,\ x_2-h_2)-2u(x_1,\ x_2)+u(x_1,\ x_2+h_2)}{h_2^2}=u_{\bar{x}_2x_2},$$

el signo $\sim$ significa la aproximación. Haciendo uso de estas expresiones, sustituyamos (1) por la ecuación en diferencias

$$\frac{y(i_1-1,\ i_2)-2y(i_1,\ i_2)+y(i_1+1,\ i_2)}{h_1^2}+$$
$$+\frac{y(i_1,\ i_2-1)-2y(i_1,\ i_2)+y(i_1,\ i_2+1)}{h_2^2}=-f(i_1,\ i_2), \quad (3)$$

o bien, en la forma abreviada,

$$y_{\bar{x}_1x_1}(i_1,\ i_2)+y_{\bar{x}_2x_2}(i_1,\ i_2)=-f(i_1,\ i_2).$$

En las designaciones sin índices tenemos

$$y_{\bar{x}_1x_1}(x)+y_{\bar{x}_2x_2}(x)=-f(x), \quad x=(i_1h_1,\ i_2h_2)\in\omega_h(G). \quad (4)$$

A esta ecuación se le debe agregar las condiciones de contorno

$$y=\mu(x), \quad x=(i_1h_1,\ i_2h_2)\in\gamma_h. \quad (5)$$

La frontera γ_h de la red está constituida por todos los nodos $(0,\ i_2)$, $(N_1,\ i_2)$, $(i_1,\ 0)$, $(i_1,\ N_2)$, a excepción de los vértices del rectángulo $(0,\ 0)$, $(0,\ N_2)$, $(N_1,\ 0)$, $(N_1,\ N_2)$ que no se emplean. La ecuación en diferencias (3) está escrita en un molde pentapuntual

$$(i_1-1,\ i_2),\ (i_1+1,\ i_2),\ (i_1,\ i_2),\ (i_1,\ i_2-1),\ (i_1,\ i_2+1).$$

El esquema (4) se denomina a menudo esquema *cruz*. Si $h_1=h_2=h$, es decir, si las redes según x_1 y x_2 coinciden, la red ω_h se llamará *cuadrada*. En esta red el esquema de diferencias (4) puede ser escrito en la forma

$$y(i_1,\ i_2)=$$
$$=\frac{y(i_1-1,\ i_2)+(y(i_1+1,\ i_2)+y(i_1,\ i_2-1)+y(i_1,\ i_2+1)+h^2f(i_1,\ i_2)}{4}.$$

Para la ecuación homogénea ($f=0$) obtenemos

$$y(i_1,\ i_2)=\frac{1}{4}[y(i_1-1,\ i_2)+y(i_1+1,\ i_2)+$$
$$+y(i_1,\ i_2-1)+y(i_1,\ i_2+1)],$$

es decir, el valor en el centro del molde se determina como media aritmética de los valores en los nodos restantes del molde.

3. Error de aproximación. Sea u = u(x) la solución del problema de Dirichlet (1), (2), y sea $y = y(i_1, i_2)$ la solución del problema en diferencias (4), (5). Veamos un error

$$z(x) = y(x) - u(x), \qquad x = (i_1h_1, i_2h_2) \in \omega_h.$$

Al sustituir $y = z + u$ en (4), (5), obtenemos para el error $z = z(x)$ una ecuación no homogénea

$$\Lambda z = z_{\bar{x}_1x_1} + z_{\bar{x}_2x_2} = -\psi(x), \quad x \in \omega_h(G), \tag{6}$$

con la condición de contorno homogénea

$$z = 0 \quad \text{cuando} \quad x \in \gamma_h. \tag{7}$$

Aquí

$$\psi(x) = \Lambda u + f(x) = u_{\bar{x}_1x_1} + u_{\bar{x}_2x_2} + f(x) \tag{8}$$

es el residuo o error de aproximación para el esquema (4) en la solución $u = u(x)$ de la ecuación (1).

Mostremos que

$$|\psi| \leqslant M_4 \frac{h_1^2 + h_2^2}{24}, \tag{9}$$

donde

$$M_4 = \max_{x \in G}\left(\left|\frac{\partial^4 u}{\partial x_1^4}\right|, \left|\frac{\partial^4 u}{\partial x_2^4}\right|\right).$$

En efecto, tomando en consideración las fórmulas

$$u(x_1 \pm h_1, x_2) = u(x_1, x_2) \pm h_1 \frac{\partial u}{\partial x_1}(y_1, x_2) + \\ + \frac{h_1^2}{2}\frac{\partial^2 u}{\partial x_1^2}(x_1, x_2) \pm \frac{h_1^3}{6}\frac{\partial^3 u}{\partial x_1^3}(x_1, x_2) + \\ + \frac{h_1^4}{24}\frac{\partial^4 u}{\partial x_1^4}(\bar{x}_1, x_2), \quad \bar{x}_1 = x_1 + \theta_1 h_1 \quad 0 \leqslant \theta_1 \leqslant 1,$$

$$u(x_1, x_2 \pm h_2) = u(x_1, x_2) \pm h_2 \frac{\partial u}{\partial x_2}(x_1, x_2) + \\ + \frac{h_2^2}{2}\frac{\partial^2 u}{\partial x_2^2}(x_1, x_2) \pm \frac{h_2^3}{6}\frac{\partial^3 u}{\partial x_2^3}(x_1, x_2) + \\ + \frac{h_2^4}{24}\frac{\partial^4 u}{\partial x_2^4}(x_1, \bar{x}_2), \quad \bar{x}_2 = x_2 + \theta_2 h_2 \quad 0 \leqslant \theta_2 \leqslant 1,$$

encontramos

$$\psi=\left(\frac{\partial^2 u}{\partial x_1^2}+\frac{\partial^2 u}{\partial x_2^2}-f(x)\right)+\frac{h_1^2}{24}\frac{\partial^4 u}{\partial x_1^4}(\bar{x}_1,\ x_2)+\frac{h_2^2}{24}\frac{\partial^4 u}{\partial x_2^4}(x_1,\bar{x}_2).$$

De aquí y de (1) proviene (9).

Así pues, el esquema (4) es de segundo orden de aproximación.

4. Esquema de orden aumentado de exactitud. Haciendo uso de un molde nonapuntual (x_1, x_2), $(x_1 \pm h_1, x_2)$, $(x_1, x_2 \pm h_2)$, $(x_1 \pm h_1, x_2 \pm h_2)$, podemos construir un esquema que tenga el cuarto orden de aproximación (y de exactitud), si suponemos que la solución del problema (1) — (2) $u = u(x) \in C^{(6)}(\bar{G})$. Dicho esquema tiene por expresión

$$\begin{gathered}\Lambda' y=\left(\Lambda_1+\Lambda_2+\frac{h_1^2+h_2^2}{12}\Lambda_1\Lambda_2\right)y=-\varphi(x),\quad x\in\omega_h,\\ y(x)=\mu(x),\quad x\in\gamma_h,\\ \Lambda_1 y=y_{\bar{x}_1x_1},\quad \Lambda_2 y=y_{\bar{x}_2x_2},\\ \varphi=f+\frac{h_1^2}{12}\Lambda_1 f+\frac{h_2^2}{12}\Lambda_2 f.\end{gathered}\qquad(10)$$

La comprobación inmediata muestra que el residuo es

$$\psi=\Lambda' u+\varphi=O(|h|^4).\qquad(11)$$

Para el error $z = y - u$, donde y es la solución del problema (10), obtenemos

$$\Lambda' z=-\psi(x),\quad x\in\omega_h;\quad z=0,\quad x\in\gamma_h.\qquad(12)$$

5. Propiedades del operador de diferencias. Sea $\mathring{y}(x)$ una función reticular definide sobre la red $\bar{\omega}_h = \omega_h(\bar{G})$ e igual a cero en la frontera γ_h de la red, y sea $\mathring{\Omega}$ un conjunto de funciones reticulares y.

El operador A se definirá del modo siguiente:

$$Ay=-\Lambda\mathring{y}=-\mathring{y}_{\bar{x}_1x_1}-\mathring{y}_{\bar{x}_2x_2}\quad\text{para cualquier } y\in\Omega,\qquad(13)$$

donde Ω es un espacio de funciones reticulares que están definidas en los nodos interiores de la red ω_h y coinciden en

dichos nodos con $\overset{\circ}{y}$, $y(x) = \overset{\circ}{y}(x)$ para $x \in \omega_h$. Al designar

$$\varphi = f + \frac{\mu(l_1, x_2)}{h_1^2} \text{ para } x_1 = l_1 - h_1, \quad 0 < x_2 < l_2,$$

$$\varphi = f + \frac{\mu(0, x_2)}{h_1^2}, \quad x_1 = h_1 \quad 0 < x_2 < l_2,$$

$$\varphi = f + \frac{\mu(x_1, l_2)}{h_2^2}, \quad 0 < x_1 < l_1, \quad x_2 = l_2 - h_2,$$

$$\varphi = f + \frac{\mu(x_1, 0)}{h_2^2}, \quad 0 < x_1 < l_1, \quad x_2 = h_2,$$

$\varphi(x) = f(x)$ en los demás puntos $x \in \omega_h$, escribamos el esquema de diferencias (4), (5) en una forma operacional

$$A\varphi = \varphi, \qquad y, \varphi \in H, \tag{14}$$

donde $H = \Omega$.

Introduzcamos en H un producto escalar

$$(y, v) = \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} \overset{\circ}{y}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 h_2$$

y probemos que el operador A es autoconjugado. Representemos A en forma de una suma $A = A_1 + A_2$, donde $A_1 y = -\overset{\circ}{y}_{\bar{x}_1 x_1}$, $A_2 y = -\overset{\circ}{y}_{\bar{x}_2 x_2}$, y mostremos que cada uno de los operadores «unidimensionales» A_1 y A_2 es autoconjugado. Será suficiente probarlo para el operador A_1. Veamos un producto escalar

$$(A_1 y, v) = -\sum_{i_2=1}^{N_2-1} h_2 \Big(\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{\bar{x}_1 x_1}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 \Big). \tag{15}$$

Aprovechemos la fórmula unidimensional de Green (cap. I, § 4):

$$\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{\bar{x}_1 x_1}(i_1, i_2) \overset{\circ}{v}(i_1, i_2) h_1 = \sum_{i_1=1}^{N_1-1} \overset{\circ}{y}(i_1, i_2) \overset{\circ}{v}_{\bar{x}_1 x_1}(i_1, i_2) h_1.$$

Sustituyendo esta expresión en (15), obtenemos

$$(A_1y, v) = -\sum_{i_2=1}^{N_2-1} h_2\Big(\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}(i_1, i_2)\,\overset{\circ}{v}_{\bar{x}_1x_1}(i_1, i_2)\,h_1\Big) = (y, A_1v).$$

De un modo análogo nos convencemos de que $A_2^* = A_2$, y, por consiguiente,

$$(Ay, v) = ((A_1 + A_2)\,y, v) = (A_1y, v) + (A_2v, y) =$$
$$= (y, A_1v) + (y, A_2v) = (y, Av),$$

es decir, $A^* = A$.

Si hacemos uso de la primera fórmula de diferencias de Green

$$\sum_{i_1=1}^{N_1-1} \overset{\circ}{y}_{\bar{x}_1x_1}(i_1, i_2)\,\overset{\circ}{y}(i_1, i_2)\,h_1 = -\sum_{i=1}^{N_1} (\overset{\circ}{y}_{\bar{x}_1}(i_1, i_2))^2h_1,$$

obtendremos

$$(A_1y, y) = \sum_{i_2=1}^{N_2-1} h_2 \sum_{i_1=1}^{N_1} (\overset{\circ}{y}_{\bar{x}_1}(i_1, i_2))^2\,h_1 > 0,$$

y, análogamente, $(A_2y, y) > 0$, de suerte que $A > 0$, es decir, A es un operador definido positivo y autoconjugado.

No es difícil encontrar las fronteras δ y Δ del operador A, es decir, los números, para los cuales se verifican las desigualdades $\delta E \leqslant A \leqslant \Delta E$, donde E es un operador unidad. En efecto, se ha mostrado en el § 4, cap. I que

$$\delta_1 \sum_{i_1=1}^{N_1-1} (\overset{\circ}{y}(i_1, i_2))^2\,h_1 \leqslant \sum_{i_1=1}^{N_1} (\overset{\circ}{y}_{\bar{x}_1}(i_1, i_2))^2h_1 \leqslant$$
$$\leqslant \Delta_1 \sum_{i_1=1}^{N_1-1} (\overset{\circ}{y}(i_1, i_2))^2h_1,$$

donde

$$\delta_1 = \frac{4}{h_1^2}\,\mathrm{sen}^2\,\frac{\pi h_1}{2l_1}, \quad \Delta_1 = \frac{4}{h_1^2}\cos^2\frac{\pi h_1}{2l_1}.$$

Al sumar estas desigualdades según $i_2 = 1, 2, \ldots, N_2 - 1$, obtendremos $\delta_1(y, y) \leqslant (A_1y, y) \leqslant \Delta_1(y, y)$.

Del modo análogo encontramos $\delta_2 (y, y) \leqslant (A_2 y, y) \leqslant \leqslant \Delta_2 (y, y)$, donde

$$\delta_2 = \frac{4}{h_2^2} \operatorname{sen}^2 \frac{\pi h_2}{2l_2}, \quad \Delta_2 = \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}.$$

De aquí se desprende

$$\delta \, \| y \|^2 \leqslant (Ay, y) \leqslant \Delta \, \| y \|^2, \tag{16}$$

donde

$$\begin{aligned} \delta &= \delta_1 + \delta_2 = \frac{4}{h_1^2} \operatorname{sen}^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \operatorname{sen}^2 \frac{\pi h_2}{2l_2}, \\ \Delta &= \Delta_1 + \Delta_2 = \frac{4}{h_1^2} \cos^2 \frac{\pi h_1}{2l_1} + \frac{4}{h_2^2} \cos^2 \frac{\pi h_2}{2l_2}, \end{aligned} \tag{17}$$

En el cuadrado ($l_1 = l_2 = 1$) en la red cuadrada ($h_1 = = h_2 = h$) tenemos

$$\delta = \frac{8}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}, \quad \Delta = \frac{8}{h^2} \cos^2 \frac{\pi h}{2}, \quad \delta + \Delta = \frac{8}{h^2}. \tag{18}$$

6. Problema de diferencias en valores propios. Planteemos un problema: hallar tales valores del parámetro λ (valores propios), para los cuales el problema homogéneo

$$y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} + \lambda y = 0, \quad x \in \omega_h, \quad y = 0, \quad x \in \gamma_h \tag{19}$$

tenga soluciones no triviales (funciones propias). Recurriremos al método de separación de variables y buscaremos la solución del problema (19) en forma de un producto

$$y(x_1, x_2) = v(x_1) \, w(x_2) \not\equiv 0 \tag{20}$$

de función $v(x_1)$, dependiente sólo de x_1, y de función $w(x_2)$ que depende sólo de x_2. Al sustituir (20) en (19) y al dividir por $y = vw$, obtendremos

$$\frac{v_{\bar{x}_1 x_1}}{v} = -\frac{w_{\bar{x}_2 x_2}}{w} - \lambda, \qquad (x, x_2) \in \omega_h. \tag{21}$$

El primer miembro depende sólo de x_1, mientras que el segundo, sólo de x_2; la igualdad (21) es posible sólo bajo la condición de que

$$\frac{v_{\bar{x}_1 x_1}}{v} = \lambda^{(1)}, \quad -\frac{w_{\bar{x}_2 x_2}}{w} - \lambda = \lambda^{(1)},$$

donde $\lambda^{(1)}$ = const. De aquí resultan dos problemas unidimensionales en valores propios para los segmentos $0 \leqslant i_1 h_1 \leqslant l_1$ y $0 \leqslant i_2 h_2 \leqslant l_2$, respectivamente:

$$v_{\bar{x}_1 x_1} + \lambda^{(1)} v = 0, \quad 0 < x_1 = i_1 h_1 < l_1,$$
$$v = 0, \quad i_1 = 0, \quad N_1, \qquad (22)$$
$$w_{\bar{x}_2 x_2} + \lambda^{(2)} w = 0, \quad 0 < x_2 = i_2 h_2 < l_2,$$
$$w = 0, \quad i_2 = 0,\ N_2, \qquad (23)$$

donde $\lambda^{(2)} = \lambda - \lambda^{(1)}$, o bien $\lambda = \lambda^{(1)} + \lambda^{(2)}$.

Recurriendo al p. 8 del § 4, cap. I, escribamos la solución de los problemas (22), (23) en la forma

$$\lambda_{k_1}^{(1)} = \frac{4}{h_1^2} \operatorname{sen}^2 \frac{\pi k_1 h_1}{2 l_1},$$
$$v_{k_1}^{(1)}(x_1) = \sqrt{\frac{2}{l_1}} \operatorname{sen} \frac{\pi k_1 x_1}{l_1}, \quad k_1 = 1, 2, \ldots, N_1 - 1,$$
$$\lambda_{k_2}^{(2)} = \frac{4}{h_2^2} \operatorname{sen}^2 \frac{\pi k_2 h_2}{2 l_2},$$
$$w_{k_2}^{(2)}(x_2) = \sqrt{\frac{2}{l_2}} \operatorname{sen} \frac{\pi k_2 x_2}{l_2}, \quad k_2 = 1, 2, \ldots, N_2 - 1,$$

donde $x_\alpha = i_\alpha h_\alpha$, $i_\alpha = 0, 1, \ldots, N_\alpha$, $\alpha = 1, 2$.

De aquí se deduce que el problema (19) tiene valores propios

$$\lambda_{k_1 k_2} = \frac{4}{h_1^2} \operatorname{sen}^2 \frac{\pi k_1 h_1}{2 l_1} + \frac{4}{h_2^2} \operatorname{sen}^2 \frac{\pi k_2 h_2}{2 l_2},$$
$$k_\alpha = 1, 2, \ldots, N_\alpha - 1, \ \alpha = 1, 2, \qquad (24)$$

y funciones propias correspondientes $y_k = v_{k_1}^{(1)}(x_1)\, w_{k_2}^{(2)}(x_2)$:

$$y_k = y_{k_1 k_2}(x_1, x_2) = \sqrt{\frac{4}{l_1 l_2}} \operatorname{sen} \frac{\pi k_1 x_1}{l_1} \operatorname{sen} \frac{\pi k_2 x_2}{l_2},$$
$$x_\alpha = i_\alpha h_\alpha, \quad i_\alpha = 0, 1, \ldots, N_\alpha, \quad k_\alpha = 1, 2, \ldots, N_\alpha - 1,$$
$$\alpha = 1, 2. \qquad (25)$$

Estas funciones propias están ortonormalizadas:

$$(y_{k_1 k_2}, y_{m_1 m_2}) = \delta_{k_1 m_1} \delta_{k_2 m_2}.$$

De (17) y (25) se ve que

$$\delta = \text{mín}\, \lambda_{k_1 k_2} = \lambda_{1,1} \quad \Delta = \text{máx}\, \lambda_{k_1 k_2} = \lambda_{N_1 - 1,\, N_2 - 1},$$

donde δ y Δ se determinan según las fórmulas (17). Para δ y Δ son justas las estimaciones

$$\delta \geqslant 8\left(\frac{1}{l_1^2}+\frac{1}{l_2^2}\right), \quad \Delta < \frac{4}{h_1^2}+\frac{4}{h_2^2}. \tag{26}$$

7. Estimación de la velocidad de convergencia del esquema «cruz». Principio del máximo. Para el error $z = y - u$ del esquema en el p. 3 se ha obtenido el problema (6), (7), donde

$$\psi(x) = O(|h|^2), \qquad |h|^2 = h_1^2 + h_2^2 \tag{27}$$

bajo el supuesto de que la solución $u = u(x) \in C^{(4)}(\overline{G})$ del problema de partida (1), (2) sea suficientemente suave. Demostremos que el esquema (4) converge con la velocidad $O(|h|^2)$ (es de segundo orden de exactitud) en la norma reticular C, es decir, que $\| z \|_C = O(|h|^2)$, donde $\| z \|_C = \max\limits_{x\in\omega_h} |z(x)|$. Para esto nos hará falta la estimación de la solución del problema (6), (7) a través del segundo miembro de ψ. El problema de contorno de Dirichlet es un caso particular del problema

$$\begin{gathered}\mathscr{L}[y] = a_{i_1i_2}y_{i_1i_2} - b_{i_1-1,\,i_2}y_{i_1-1,\,i_2} - b_{i_1+1,\,i_2}y_{i_1+1,\,i_2} - \\ - b_{i_1,\,i_2-1}y_{i_1,\,i_2-1} - b_{i_1,\,i_2+1}y_{i_1,\,i_2+1} = \varphi_{i_1i_2}, \\ x = (i_1h_1,\ i_2h_2) \in \omega_h; \quad y = \mu, \quad x \in \gamma_h,\end{gathered} \tag{28}$$

donde $a = a_{i_1i_2}$, $b = b_{i_1,\,i_2}$ son los coeficientes.
En el caso (4) tenemos

$$\begin{gathered} a_{i_1i_2} = 2\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right), \\ b_{i_1\pm1,\,i_2} = \frac{1}{h_1^2}\,b_{i_1,\,i_2\pm1} = \frac{1}{h_2^2}, \\ b_{i_1,\,i_2} = 0. \end{gathered} \tag{29}$$

El operador $\mathscr{L}[y]$ puede anotarse de otra forma:

$$\begin{gathered}\mathscr{L}[y] = d_{i_1i_2}y_{i_1i_2} + b_{i_1-1,\,i_2}(y_{i_1,\,i_2} - y_{i_1-1,\,i_2}) + \\ + b_{i_1+1,\,i_2}(y_{i_1,\,i_2} - y_{i_1+1,\,i_2}) + b_{i_1,\,i_2-1}(y_{i_1,\,i_2} - y_{i_1,\,i_2-1}) + \\ + b_{i_1i_2+1}(y_{i_1i_2} - y_{i_1,\,i_2+1}),\end{gathered} \tag{30}$$

donde $d_{i_1i_2} = a_{i_1i_2} - b_{i_1-1,\,i_2} - b_{i_1+1,\,i_2} - b_{i_1,\,i_2-1} - b_{i_1,\,i_2+1}$.

Supondremos cumplidas las condiciones

$$d = d_{i_1 i_2} \geqslant 0, \quad b_{i_1 \pm 1, i_2} > 0, \quad b_{i_1, i_2 \pm 1} \geqslant 0. \tag{31}$$

Para el problema (4) tenemos $d \equiv 0$.

TEOREMA 1. *Supongamos cumplidas las condiciones* (31) *y que* $\varphi(x) \geqslant 0$, $y|_\gamma \geqslant 0$. *Entonces, la solución de la ecuación* (28) *es no negativa, es decir,* $y(x) \geqslant 0$ *en todos los nodos de la red* $\bar{\omega}_h = \omega_h(\bar{G})$.

DEMOSTRACIÓN. Supongamos que la afirmación del teorema es falsa y existe por lo menos un nodo $x_{i_*} = (i_1^0 h_1, i_2^0 h_2)$ en el cual $y(x_{i_0}) < 0$. Entonces la función $y(x)$ ha de tomar en cierto nodo interior de la red un valor negativo mínimo $\min\limits_{x \in \omega_h} y(x) = y(x_*)$. En este nodo se verifica la ecuación (28). Si $d(x_*) = 0$ y $\varphi(x_*) = 0$, la ecuación (28) se verificará sólo bajo la condición de que $y(x) = y(x_*)$ en todos los nodos del molde. Sin embargo, por cuanto $\varphi(x) \not\equiv 0$, existe un nodo $x_{i_{**}}$ en el que $y(x_{i_{**}}) = y(x_{i_*}) = \min y(x) = c_0 < 0$, y, al menos en un nodo, por ejemplo, para $x = x_{i_1+1}$ tenemos $y_{i_1+1} > c_0$, y, por consiguiente, $\mathscr{L}[y]|_{x = x_{i_{**}}} < 0$, que contradice la condición $\mathscr{L}[y] = \varphi(x) \geqslant 0$. La contradicción obtenida demuestra el teorema.

TEOREMA 2 (TEOREMA DE COMPARACIÓN). *Sea* $\bar{y}(x)$ *la solución del problema*

$$\mathscr{L}[\bar{y}] = \bar{\varphi}, \quad x \in \omega_h, \quad \bar{y} = \bar{\mu}, \quad x \in \gamma_h \tag{32}$$

y supongamos cumplidas las condiciones (31). *Si*

$$|\varphi(x)| \leqslant \bar{\varphi}(x), \quad x \in \omega_h, \quad |\mu(x)| \leqslant \bar{\mu}(x), \quad x \in \gamma_h, \tag{33}$$

entonces, para la ecuación del problema (28) *es justa la estimación*

$$|y(x) \leqslant \bar{y}(x) \ \textit{para todos los} \ x \in \bar{\omega}_h.$$

Basta por convencerse de que para las funciones $u = \bar{y}(x) + y(x)$, $v = \bar{y}(x) - y(x)$ quedan cumplidas las condiciones del teorema 1, y, por lo tanto, $u(x) \geqslant 0$, $v(x) \geqslant 0$, o bien $y(x) \geqslant -\bar{y}(x)$, $y(x) \leqslant \bar{y}(x)$, es decir, $|y| \leqslant \bar{y}$.

Así pues, la función $\bar{y}(x)$ es una *mayorante*. Si la mayorante $\bar{y}(x)$ queda hallada, la solución del problema (28) viene estimada de acuerdo con el teorema 2. Para el problema (4) elijamos, a título de mayorante, una función

$$\bar{y}(x) = C[L^2 - (x_1^2 + x_2^2)], \qquad L^2 = l_1^2 + l_2^2. \quad (34)$$

Calculemos al principio $\bar{\varphi} = \mathscr{L}[\bar{y}] = -\Lambda\bar{y} = C\Lambda(x_1^2 + x_2^2) = C(\Lambda_1 x_1^2 + \Lambda_2 x_2^2) = 4C$, puesto que $(x_1^2)_{\bar{x}_1 x_1} = \frac{1}{h_1^2}((x_1 + h_1)^2 - 2x_1^2 + (x_1 - h_1)^2) = 2$. De la fórmula (34) se ve que $\bar{\mu} = y(x) > 0$ en la frontera γ_h. Volvamos ahora al problema (6), (7) para el error $z = y - u$ del esquema (4). Eligiendo $4C = \| \psi \|_C$, y teniendo presente que $z|_{\gamma_h} = 0$, obtenemos $| z(x) | < \bar{y}(x) < CL^2$, de suerte que

$$\| z \|_C \leqslant \frac{L^2}{4} \| \psi \|_C. \quad (35)$$

De aquí y de (9) se deduce la convergencia uniforme del esquema (4) con el segundo orden de exactitud.

OBSERVACION. La ecuación (28) puede ser sustituida por una ecuación de la forma más general

$$\mathscr{L}[y] = a(x)\,y(x) - \sum_{\substack{\xi \in \sigma(x) \\ \xi \neq x}} b(x, \xi)\,y(\xi) = \varphi(x), \quad (36)$$

donde $a(x) > 0$, $b(x, \xi) > 0$, $\sigma(x)$ es un conjunto de nodos $\xi \neq x$ del molde con centro en el nodo x, con la particularidad de que

$$d(x) = a(x) - \sum_{\xi \in \sigma(x)} b(x, \xi) \geqslant 0.$$

Para la ecuación (36) son verídicos los teoremas 1 y 2. Cuando se trata de un esquema con el orden aumentado de exactitud, el molde consta de nueve nodos, el conjunto $\sigma(x)$ de ocho nodos, y en este caso $a = \frac{5}{3}(h_1^{-2} + h_2^{-2})$, mientras que en el segundo miembro se tienen coeficientes $\frac{1}{6}(5h_1^{-2} - h_2^{-2})$,

$\frac{1}{6}(5h^{-2}-h^{-2})$, los cuales son positivos sólo a condición de que

$$1/\sqrt{5} \leqslant h_1/h_2 \leqslant \sqrt{5},$$

y, por consiguiente, la estimación (35) se obtendrá bajo dicha condición.

§ 2. Resolución de las ecuaciones en diferencias

1. Métodos directos. Método de separación de variables. El sistema de ecuaciones en diferencias para el problema de Dirichlet del § 1,

$$\Lambda y = y_{\bar{x}_1 x_1} + y_{\bar{x}_2 x_2} = -f(x), \quad x \in \omega_h \quad y = \mu, \quad x \in \gamma_h \quad (1)$$

cuenta con una matriz de alto orden $(N_1-1)(N_2-1)$. Se toman habitualmente N_1, $N_2 \sim 50-100$, de modo que el número de ecuaciones en el sistema (1) es igual a 10^3—10^4. La resolución de un sistema de orden tan alto, por el método de Gauss, exigiría aproximadamente $(N_1-1)^3(N_2-1)^3$, esto es, 10^9—10^{12} operaciones, si el sistema (1) no tuviera una calidad muy buena: la matriz del sistema está débilmente llenada y sólo tiene $\sim 5N_1N_2$ elementos distintos de cero. Por esta razón, para la resolución de un sistema de ecuaciones en diferencias se logra construir los métodos que requieren $O(N \ln N)$ e incluso $O(N)$ operaciones, donde $N=(N_1-1)(N_2-1)$. Describamos uno de los métodos directos de resolución del problema en diferencias de Dirichlet de la ecuación de Poisson en un rectángulo.

Escribamos el problema (1) en una forma

$$\Lambda \dot{y} = \dot{y}_{\bar{x}_1 x_1} + \dot{y}_{\bar{x}_2 x_2} = -\varphi(x), \quad x \in \omega_h, \quad \dot{y}\,|_{\gamma_h} = 0, \quad (2)$$

donde $\dot{y}(x) \equiv y(x)$ para $x \in \omega_h$, y $\varphi(x)$ se determina según las fórmulas (14) del § 1.

Su solución puede encontrarse por el método de separación de variables. Sean $\{v_{k_2}^{(2)}(x_2), \lambda_{k_2}^{(2)}\}$ $(k=1, 2, \ldots, \ldots, N_2-1)$ funciones propias y valores propios del problema

$$\Lambda_2 v + \lambda v = 0, \qquad x \in \omega_h; \quad v(0) = v(l_2) = 0. \quad (3)$$

Las expresiones para $\lambda_{k_2}^{(2)}$ y $v_{k_2}^{(2)}(x_2)$ se han aducido en el p. 6, § 1.

Desarrollemos la solución $\mathring{y}(x_1, x_2)$ y el segundo miembro $\varphi(x_1, x_2)$ según las funciones propias $\{v_{k_2}^{(2)}\}$:

$$\mathring{y}(x_1, x_2) = \sum_{k_2=1}^{N_2-1} c_{k_2}(x_1)\, v_{k_2}(x_2), \tag{4}$$

$$\varphi(x_1, x_2) = \sum_{k_2=1}^{N_2-1} \varphi_{k_2}(x_1)\, v_{k_2}(x_2), \tag{5}$$

donde $x_\alpha = i_\alpha h_\alpha$, $i_\alpha = 1, 2, \ldots, N_\alpha - 1$, $\alpha = 1, 2$, $c_{k_2}(x_1)$ y $\varphi_{k_2}(x_1)$ son los coeficientes de Fourier, por ejemplo,

$$\varphi_{k_2}(x_1) = \sum_{i_2=1}^{N_2-1} h_2 \varphi(x_1, i_2h_2)\, v_{k_2}(i_2h_2).$$

Apliquemos un operador $\Lambda = \Lambda_1 + \Lambda_2$ al producto $c_{k_2}v_{k_2}$:

$$\Lambda c_{k_2}(x_1)\, v_{k_2}(x_2) =$$
$$= v_{k_2}(x_2)\, \Lambda_1 c_{k_2}(x_1) + c_{k_2}(x_1)\, \Lambda_2 v_{k_2}(x_2) =$$
$$= v_{k_2}(x_2)\, \Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1)\, v_{k_2}(x_2) =$$
$$= [\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1)]\, v_{k_2}(x_2).$$

Ahora, al sustituir esta expresión en (2) y al tomar en consideración (5), obtendremos

$$\sum_{k_1=1}^{N_2-1} \{\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) + \varphi_{k_2}(x_1)\}\, v_{k_2}(x_2) = 0. \tag{6}$$

En virtud de que la función $\{v_{k_2}(x_2)\}$ es ortogonal, esta identidad se verifica sólo cuando es igual a cero la expresión encerrada dentro de las llaves

$$\Lambda_1 c_{k_2}(x_1) - \lambda_{k_2}^{(2)} c_{k_2}(x_1) = -\varphi_{k_2}(x_1),$$
$$k_2 = 1, 2, \ldots, N_2 - 1,$$
$$= i_1h_1,\ 0 < i_1 < N_1,\ c_{k_2}(i_1h_1) = 0,\ i_1 = 0, N_1. \tag{7}$$

Efectivamente, multiplicando (6) escalarmente por $v_{k_2}(x_2)$, tenemos

$$0=\sum_{k=1}^{N_2-1}\{\cdot\}_k(v_k, v_{k_2})=\sum_{k=1}^{N_2-1}\{\cdot\}_k\delta_{kk_2}=\{\cdot\}_{k_2}=0,$$

donde $\{\cdot\}_{k_2}$ es el contenido de las llaves (6).

Los problemas (7) se resuelven por el método de factorización; se necesita emplear N_2-1 veces en total el alogritmo de factorización para $k_2=1, 2, \ldots, N_2-1$. Conociendo $c_{k_2}(x_1)$, hallaremos la solución del problema (2) según la fórmula (4). Con este fin se requiere, primero, calcular los coeficientes de Fourier $\varphi_{k_2}(x_1)$ $(k_2=1, 2, \ldots, N_2-1)$. De las fórmulas (4) y (5) se ve que $y(x_1, x_2)$ y $\varphi_{k_2}(x_1)$ se calculan según las fórmulas de una misma forma:

$$w_i=\sum_{k=1}^{N-1}\alpha_k \operatorname{sen}\frac{k\pi i}{N}, \quad i=1, 2, \ldots, N-1. \qquad (8)$$

Se ha elaborado un algoritmo especial de transformación rápida de Fourier para calcular sumas, el cual permite obtener la suma (8) en el transcurso de $5N\log_2 N$ operaciones aritméticas (cuando $N=2^n$, siendo n un número entero) en lugar de $O(N^2)$ si se usa el modo de sumar habitual. Este algoritmo permite hallar la solución del problema de partida (2) en el transcurso de $O(N_1N_2\log_2 N_2)$ operaciones. El método de separación de variables puede combinarse con el de reducción o descomposición que representa una modificación del método de Gauss. De resultas obtenemos un algoritmo con el número de operaciones $Q\approx 5N_1N_2\log_2 N_2$ que es dos veces menor que el número correspondiente para el algoritmo de separación aducido más arriba.

2. Métodos iterativos. Los métodos directos son más económicos cuando se resuelve el problema de Dirichlet en diferencias para la ecuación de Poisson en un rectángulo. Actualmente existen programas estándar en el lenguaje algorítmico FORTRAN y ALGOL para resolver las ecuaciones de Poisson en un rectángulo con las condiciones de contorno de tres tipos y también con las condiciones de contorno mixtas. No obstante, cuando el dominio no es un rectángulo o cuando se analizan las ecuaciones de coeficientes variables,

se aplican los métodos iterativos. En realidad los métodos directos son económicos sólo en el caso en que las variables se separen.

En el cap. III se ha estudiado la teoría de los métodos iterativos para una ecuación.

$$Ay = \varphi,$$

donde $A = A^* > 0$. La comparación de diferentes métodos se realizaba para el problema unidimensional modelo en el segmento $0 \leqslant x \leqslant 1$:

$$y_{\bar{x}x} = -f(x), \quad x = ih, \quad 0 < i < N, \quad y_0 = y_N = 0.$$

Para el problema citado el operador A tiene la forma $Ay = -\overset{\circ}{y}_{\bar{x}x}$. Las fronteras del operador se determinan por las constantes

$$\delta = \frac{4}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}, \qquad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}.$$

El número de iteraciones para los métodos, estudiados en el cap. III depende de la razón

$$\eta = \frac{\delta}{\Delta} = \operatorname{tg}^2 \frac{\pi h}{2} \approx \frac{\pi^2 h^2}{4}. \tag{9}$$

Veamos ahora, a título del problema modelo, un problema de Dirichlet bidimensional en un cuadrado unitario ($l_1 = l_2 = 1$) sobre la red cuadrada de paso $h = h_1 = h_2$:

$$Ay = -\overset{\circ}{y}_{\bar{x}_1 x_1} - \overset{\circ}{y}_{\bar{x}_2 x_2} = \varphi, \quad \varphi,\ y \in H. \tag{10}$$

El número de intervalos por cada una de las direcciones es N, de modo que $h = 1/N$.

Las fronteras δ y Δ del operador A se han determinado en el § 1 (véase (18) del § 1), la razón $\eta = \delta/\Delta$ coincide con (9). De aquí proviene que el número de iteraciones no depende del número de mediciones (si $h_1 \neq h_2$, $l_1 \neq l_2$, depende poco). Por esta razón, las estimaciones del número de iteraciones de diferentes métodos iterativos, obtenidas para el problema modelo unidimensional, siguen siendo en rigor para el caso bidimensional.

En el caso de una red no cuadrada, el número de iteraciones para el problema bidimensional puede diferir un poco

del número de iteraciones para el problema unidimensional.

Aquí se examinará sólo el método iterativo alternado triangular para resolver el problema de Dirichlet en diferencias (10).

3. Método alternado triangular. Para la resolución de una ecuación operacional

$$Au = f, \quad A = A^* > 0, \quad A\colon H \to H, \tag{11}$$

hemos considerado en el cap. III los métodos iterativos de un paso (de dos capas), los cuales se anotaban en la siguiente forma canónica:

$$B \frac{y_{k+1} - y_k}{\tau_{k+1}} + Ay_k = f, \quad k = 0, 1, \ldots, n, \quad \text{para todo } y_0 \in H, \tag{12}$$

donde $B\colon H \to H$, $B = B^* > 0$. Para A y B se cumplen las condiciones

$$\gamma_1 B \leqslant A \leqslant \gamma_2 B, \qquad \gamma_1 > 0, \tag{13}$$

donde γ_1 y γ_2 son unas constantes.

El número mínimo de iteraciones $\min_{\tau_k} n(\varepsilon)$ con γ_1 y γ_2 prefijadas se consigue al elegir los parámetros de Chébishev

$$\tau_k = \frac{\tau_0}{1 + \rho_0 \sigma_k}, \quad \tau_0 = \frac{2}{\gamma_1 + \gamma_2} \quad \rho_0 = \frac{1 - \xi}{1 + \xi}, \quad \xi = \frac{\gamma_1}{\gamma_2}, \qquad k = 1, 2, \ldots, n, \tag{14}$$

donde σ_k pertenece a cierto conjunto, especialmente ordenado, de ceros del polinomio de Chébishev; con tal ordenación el método (12) es estable desde el punto de vista de los cálculos.

Para determinar la $(k + 1)$-ésima iteración tenemos una ecuación

$$By_{k+1} = F_k, \qquad F_k = By_k - \tau_{k+1}(Ay_k - f).$$

El número de operaciones al calcular y_{k+1} depende de B. Al elegir

$$B = (D + \omega A_1) D^{-1} (D + \omega A_2), \tag{15}$$

donde A_1 y A_2 son los operadores con matrices triangulares $A_1^* = A_2$, $A_1 + A_2 = A$ y $D = D^* > 0$ es un operador

arbitrario, obtenemos el método alternado triangular. Corrientemente, $D = (d_{i_1 i_1})$ es una matriz diagonal. En el cap. III fue elaborada la teoría de este método y se han encontrado las constantes γ_1, γ_2 y ω para las condiciones prefijadas

$$A \geqslant \delta D, \quad A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A, \quad \delta > 0, \quad \Delta \geqslant \delta > 0, \qquad (16)$$

las cuales pueden ser escritas en la forma equivalente:

$$(Ay, y) \geqslant \delta (Dy, y), \quad (D^{-1} A_2 y, A_2 y) \leqslant \frac{\Delta}{4} (Ay, y).$$

En este caso tenemos

$$\omega = \frac{2}{\sqrt{\delta\Delta}}, \qquad \xi = \frac{2\sqrt{\eta}}{1+\sqrt{\eta}}, \qquad \eta = \frac{\delta}{\Delta}, \qquad (17)$$

y para el número de iteraciones es justa la estimación

$$n(\varepsilon) \approx n_0(\varepsilon) = \frac{1}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{3}. \qquad (18)$$

4. Método alternado triangular para el problema de Dirichlet en diferencias. Volvamos al problema (10). Representemos el operador A en forma de una suma $A = A_1 + A_2$, donde

$$A_1 y = \frac{y_{\bar{x}_1}}{h_1} + \frac{y_{\bar{x}_2}}{h_2}, \qquad A_2 y = -\frac{y_{x_1}}{h_1} - \frac{y_{x_2}}{h_2},$$

y pongamos $D = E$. El carácter conjugado de A_1 y A_2: $A_2 = A_1^*$ se establece por comparación de sus matrices o bien con ayuda de la primera fórmula de Green en diferencias: $(A_1 y, v) = (y, A_1^* v) = (y, A_2 v)$.

Para determinar y_{k+1} obtenemos una ecuación

$$By_{k+1} = (E + \omega A_1)(E + \omega A_2) y_{k+1} = F_k,$$

$$F_k = B\overset{\circ}{y}_k + \tau_{k+1} (\Lambda y_k + \varphi) \ (y_k = \mu, \ \overset{\bullet}{y}_k = 0 \text{ para } x \in \gamma_h).$$

Los valores de y_{k+1} se hallan sucesivamente de la ecuación

$$(E + \omega A_1) \overset{\bullet}{y}{}_k^{(1)} = F_k, \qquad (E + \omega A_2) \overset{\bullet}{y}_{k+1} = \overset{\bullet}{y}{}_k^{(1)}.$$

De aquí obtenemos las fórmulas

$$\overset{\bullet}{y}_k^{(1)}(i_1, i_2) = \left[\frac{\varkappa_1 \overset{\bullet}{y}_k^{(1)}(i_1-1, i_2) + \varkappa_2 \overset{\bullet}{y}_k^{(1)}(i_1, i_2-1) + F_k(i_1, i_2)}{(1+\varkappa_1+\varkappa_2)}\right],$$

$$\varkappa_1 = \frac{\omega}{h_1^2}, \qquad \varkappa_2 = \frac{\omega}{h_2^2},$$

$$\overset{\bullet}{y}_{k+1}(i_1, i_2) = \left[\frac{\varkappa_1 \overset{\bullet}{y}_{k+1}(i_1+1, i_2) + \varkappa_2 \overset{\bullet}{y}_{k+1}(i_1, i_2+1) + \overset{\bullet}{y}_k^{(1)}(i_1, i_2)}{(1+\varkappa_1+\varkappa_2)}\right]. \tag{19}$$

Para determinar $\overset{\bullet}{y}_k^{(1)}(i_1, i_2)$ elijamos un nodo $i_1 = 1$, $i_2 = 1$ en la esquina izquierda del rectángulo; entonces, los dos nodos restantes $(i_1 - 1, i_2)$ y $(i_1, i_2 - 1)$ del molde $\{(i_1, i_2), (i_1 - 1, i_2) (i_1, i_{2-1})\}$ se disponen en la frontera y, por lo tanto, $\overset{\bullet}{y}^{(1)}(i_1 - 1, i_2) = \overset{\bullet}{y}^{(1)}(i_1, i_2 - 1) = 0$ son conocidos. Sabiendo $\overset{\bullet}{y}_k^{(1)}$ para $i_1 = 1$, $i_2 = 1$, hallamos sucesivamente $\overset{\bullet}{y}_k^{(1)}$ para $i_1 = 2, 3, \ldots, N_1 - 1$ y $i_2 = 1$ (en la primera fila). Luego suponemos $i_2 = 2$ y encontramos sucesivamente $\overset{\bullet}{y}_k^{(1)}$ en la segunda fila para $i_1 = 1, 2, \ldots$ $\ldots, N - 1$. Con el fin de hallar $\overset{\bullet}{y}_{k+1}$ realizamos los cálculos en el molde $\{(i_1, i_2), (i_1 + 1, i_2), (i_1, i_2 + 1)\}$ según las columnas de arriba abajo: fijamos $i_1 = N_1 - 1, N_1 -$ $- 2, \ldots, 2, 1$, y para cada i_1 cambiamos $i_2 = N_2 - 1$, $N_2 - 2, \ldots, 2, 1$. Empezamos la cuenta de $\overset{\bullet}{y}_{k+1}$ con el nodo $(i_1 = N_1 - 1, i_2 = N_2 - 1)$ en la esquina derecha superior. Se debe observar que la cuenta de $\overset{\bullet}{y}_{k+1}$ puede realizarse también según las filas de derecha a izquierda: fijamos $i_2 = N_2 - 1, N_2 - 2, \ldots, 2, 1$ y para cada i_2 cambiamos $i_1 = N_1 - 1, N_1 - 2, \ldots, 2, 1$. Es más, el cálculo de $\overset{\bullet}{y}_k^{(1)}$ podemos llevarlo a cabo no por las filas, sino por las columnas de abajo arriba. Esto lo muestran las mismas fórmulas.

Los cálculos se realizan por las fórmulas recurrentes (19); la cuenta es, evidentemente, estable. El algoritmo del tipo semejante se llama (como ya se ha indicado) *algoritmo de cómputo móvil*.

Calculemos el número de operaciones aritméticas que corresponden a un nodo de la red: el cálculo de F_k requiere

10 operaciones de adición y 10 operaciones de multiplicación; el cálculo de y_{k+1} con F_k prefijada exige 4 operaciones de adición y 6 operaciones de multiplicación.

En total se exigen 14 operaciones de adición y 16 operaciones de multiplicación para determinar y_{k+1} en un solo nodo. El número de operaciones puede ser disminuido conservando en la memoria de acceso rápido no una, sino dos sucesiones, $\{y_k\}$ y $\{w_{k+1}\}$, y empleando para la determinación de y_{k+1} un algoritmo

$$(E + \omega A_1)\, \dot{w}_{k+1/2} = \Lambda y_k + f,\ (E + \omega A_2)\dot{w}_{k+1} = \dot{w}_{k+1/2},$$
$$y_{k+1} = y_k + \tau_{k+1}\dot{w}_{k+1}.$$

En este caso para pasar de y_k a y_{k+1} son suficientes 10 operaciones de adición y 10 operaciones de multiplicación por un nodo.

5. Elección de los parámetros del método alternado triangular para el problema de Dirichlet en diferencias. Para poder aprovechar la teoría general expuesta en el cap. III (véase el § 5, cap. III), es necesario hallar las constantes δ y Δ que intervienen en la condición (16). En nuestro caso $A = A_1 + A_2 \geqslant \delta E$, donde δ es el valor propio mínimo del operador A igual a

$$\delta = 4\left(\frac{1}{h_1^2}\operatorname{sen}^2\frac{\pi h_1}{2l_1} + \frac{1}{h_2^2}\operatorname{sen}^2\frac{\pi h_2}{2l_2}\right). \tag{20}$$

Examinemos el operador $A_1 D^{-1} A_2 = A_1 A_2$. Teniendo presente que

$$A_1^* = A_2,\ (a_1b_1 + a_2b_2)^2 \leqslant (a_1^2 + a_2^2)\,(b_1^2 + b_2^2),$$

encontramos

$$(A_1A_2y,\ y) = (A_2y,\ A_2y) =$$
$$= \left(\left(\frac{1}{h_1} y_{\bar{x}_1} + \frac{1}{h_2} y_{\bar{x}_2}\right)^2,\ 1\right) \leqslant \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)((y_{\bar{x}_1})^2 + (y_{\bar{x}_2})^2,\ 1)) =$$
$$= \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right) \sum_{i_1=1}^{N_1-1} \sum_{i_2=1}^{N_2-1} [(y_{\bar{x}_1})^2 + (y_{\bar{x}_2})^2]_{i_1,i_2} h_1h_2 \leqslant$$
$$\leqslant \left(\frac{1}{h_1^2} + \frac{1}{h_2^2}\right)(Ay,\ y),$$

puesto que (véase el § 1, cap. V)

$$(Ay, y) = \sum_{i_2=1}^{N_2-1} h_2 \sum_{i_1=0}^{N_1-1} (y_{\bar{x}_1})^2_{i_1 i_2} h_1 + \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=0}^{N_2-1} (y_{\bar{x}_2})^2_{i_1 i_2} h_2.$$

Al comparar las desigualdades

$$(A_1A_2y, y) \leqslant \left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right)(Ay, y) \quad \text{y} \quad A_1A_2 \leqslant \frac{\Delta}{4} A,$$

concluimos que

$$\Delta = 4\left(\frac{1}{h_1^2}+\frac{1}{h_2^2}\right). \tag{21}$$

Conociendo δ y Δ, encontramos $\eta = \delta/\Delta$, y, según las fórmulas del § 5 del cap. V, determinamos los parámetros γ_1, γ_2, ξ, después de lo cual estimamos el número de iteraciones por la fórmula

$$n(\varepsilon) \approx \ln\frac{\varepsilon}{2}\Big/\ln\frac{1}{\rho_1}, \quad \rho_1 = \frac{1-\sqrt{\xi}}{1+\sqrt{\xi}}.$$

Haciendo uso de $n(\varepsilon)$, eligimos una totalidad estable de los parámetros de Chébishev σ_k, τ_{k+1} y $\omega = 2/\sqrt{\delta\Delta}$.

Comparamos los siguientes métodos de resolución referente al número de iteraciones $n_0(\varepsilon)$: el método de iteración simple ($n_0^{(1)}(\varepsilon)$), el esquema explícito con una totalidad de Chébishev ($n_0^{(2)}(\varepsilon)$) y el método alternado triangular ($n_0^{(3)}(\varepsilon)$) para el problema bidimensional modelo (10), empleando las fórmulas aproximadas $n_0^{(1)}(\varepsilon) \approx 2/h^2$, $n_0^{(2)}(\varepsilon) \approx \approx 3{,}2/h$, $n_0^{(3)}(\varepsilon) \approx 2{,}9\sqrt{h}$ para $\varepsilon = 10^{-4}$ (tabla 2).

TABLA 2

h	$n_0^{(1)}(\varepsilon)$	$n_0^{(2)}(\varepsilon)$	$n_0^{(3)}(\varepsilon)$
1/10	200	32	9
1/50	5 000	160	21
1/100	20 000	320	29

6. Ecuaciones en diferencias con coeficientes variables. Supongamos que se pide resolver en un rectángulo $\bar{G} = \{(x_1, x_2): 0 \leqslant x_\alpha \leqslant l_\alpha, \alpha = 1, 2\}$ un problema de Dirichlet

para la ecuación elíptica de coeficientes variables:

$$Lu = L_1u + L_2u = -f(x),$$
$$x = (x_1, x_2) \in G, \quad u = \mu(x), \quad x \in \Gamma, \quad (22)$$
$$L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha(x)\frac{\partial u}{\partial x_\alpha}\right), \quad 0 < c_1 \leqslant k_\alpha(x) \leqslant c_2, \quad \alpha = 1, 2,$$

donde c_1 y c_2 son unas constantes. Para $k_1 \equiv k_2 \equiv 1$ obtenemos la ecuación de Poisson $\Delta u = -f$.

El esquema de diferencias se construye sobre una red $\omega_h = \{x_i = (i_1h_1, i_2h_2) \mid i_\alpha = 0, 1, \ldots, N_\alpha, h_\alpha = l_\alpha/N_\alpha, \alpha = 1, 2\}$. Todo operador L_α se sustituye en el molde tripuntual $(x_\alpha - h_\alpha, x_\alpha + h_\alpha)$ por un operador de diferencias

$$\Lambda_\alpha u = (a_\alpha u_{\bar{x}_\alpha})_{x_\alpha} = \frac{1}{h_\alpha}\left[\frac{a_\alpha^{(+1_\alpha)}(u^{(+1_\alpha)} - u)}{h_\alpha} - \frac{a_\alpha(u - u^{(-1_\alpha)})}{h_\alpha}\right],$$

donde $u^{(\pm 1_1)} = u((i_1 \pm 1)h_1, i_2h_2)$, $u^{(\pm 1_2)} = u(i_1h_1, (i_2 \pm \pm 1)h_2)$. Para a_1 y a_2 pueden elegirse las expresiones más simples

$$a_1(x_1, x_2) = k_1(x_1 - 1/2h_1, x_2) = k^{(-1/2_1)},$$
$$a_2(x_1, x_2) = k_2(x_1, x_2 - 1/2h_2) = k^{(-1/2_2)},$$

que aseguran el segundo orden de aproximación:

$$\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2).$$

De resultas, al operador Lu se le pone en correspondencia un operador de diferencias sobre el molde pentapuntual:

$$\Lambda u = \Lambda_1 u + \Lambda_2 u = (a_1 u_{\bar{x}_1})_{x_1} + (a_2 u_{\bar{x}_2})_{x_2}.$$

Escribamos un esquema de diferencias

$$\Lambda y = -f(x), \; x \in \omega_h, \; y = \mu(x), \; x \in \gamma_h, \quad (23)$$
$$0 < c_1 \leqslant a_\alpha \leqslant c_2, \quad \alpha = 1, 2,$$

correspondiente al problema (22).

Introduzcamos en un espacio de funciones reticulares $H = \Omega_N$ el operador

$$Ay = -\Lambda\overset{\circ}{y}, \qquad A = A_1 + A_2,$$
$$A_1 y = -\Lambda_1\overset{\circ}{y}, \qquad A_2 y = -\Lambda_2\overset{\circ}{y}$$

y escribamos (23) en la forma operacional

$$Ay = \varphi, \qquad y,\ \varphi \in H,$$

donde φ difiere de f sólo en 4 nodos de frontera $(i_1 = 1,\ N_1 - 1,\ 0 < i_2 < N_2)$ y $(0 < i_1 < N_1,\ i_2 = 1,\ N_2 - 1)$.

El operador A es, evidentemente, autoconjugado: $(Ay,\ v) = (y,\ Av)$.

De la fórmula

$$-\sum_{i_1=1}^{N_1-1} (a_1 \mathring{y}_{\bar{x}_1})_{x_1,\, i_1} \mathring{y}_{i_1} h_1 = \sum_{i_1=1}^{N_1} (a_1 (\mathring{y}_{\bar{x}_1})^2)_{i_1} h_1$$

y de la desigualdad $0 < c_1 \leqslant a \leqslant c_2$ proviene que

$$c_1 (Ry,\ y) \leqslant (Ay,\ y) \leqslant c_2 (Ry,\ y), \text{ o bien } c_1 R \leqslant A \leqslant c_2 R, \qquad (24)$$

donde R es el operador de Laplace estudiado más arriba

$$Ry = -\dot{y}_{\bar{x}_1 x_1} - \dot{y}_{\bar{x}_2 x_2}. \qquad (25)$$

De aquí concluimos que

$$c_1 \dot{\delta} E \leqslant A \leqslant c_2 \dot{\Delta} E,$$

donde $\dot{\delta}$ y $\dot{\Delta}$ se definen por las fórmulas (20), (21).

Para resolver el problema (23) podemos aprovechar el método alternado triangular con un operador

$$B = (E + \omega R_1)(E + \omega R_2), \quad R_1 + R_2 = R, \quad R_1^* = R_2 \text{ para } D = E.$$

En este caso tenemos $\gamma_1 B \leqslant A \leqslant \gamma_2 B$, donde $\gamma_1 = c_1 \dot{\gamma}_1$, $\gamma_2 = c_2 \dot{\gamma}_2$, mientras que $\dot{\gamma}_1$ y $\dot{\gamma}_2$ se han encontrado para el operador (25). Para el número de iteraciones tenemos la siguiente estimación

$$n_0(\varepsilon) \approx \sqrt{\frac{c_2}{c_1}}\, \bar{n}_0(\varepsilon), \qquad \bar{n}_0(\varepsilon) = \frac{2}{2\sqrt{2}\sqrt[4]{\eta}} \ln \frac{2}{\varepsilon}.$$

Para una ecuación de coeficientes variables se requieren $\sqrt{c_2/c_1}$ veces más iteraciones que para la ecuación de Poisson.
Sin embargo, podemos omitir la introducción del operador R, correspondiente al operador de Laplace, representando inmediatamente el operador de coeficientes variables en la forma

$$A = A_1 + A_2,$$

$$A_1 y = \frac{1}{h_1}\left(a_1 y_{\bar{x}_1} + \frac{1}{2} y a_{x_1}\right) + \frac{1}{h_2}\left(a_2 y_{\bar{x}_2} + \frac{1}{2} y a_{x_2}\right),$$

$$A_2 y = -\frac{1}{h_1}\left(a_1^{(+1_1)} y_{x_1} + \frac{1}{2} y a_{x_1}\right) - \frac{1}{h_2}\left(a_2^{(+1_2)} y_{x_2} + \frac{1}{2} y a_{x_2}\right).$$

El operador B se elige en la forma

$$B = (D + \omega A_1)\, D^{-1}\, (D + \omega A_2), \qquad (26)$$

donde $D = d(x)\, E$ es una matriz diagonal. Para poder aplicar la teoría general se deben hallar las constantes δ y Δ que figuran en las condiciones $A \geqslant \delta D$, $A_1 D^{-1} A_2 \leqslant \frac{\Delta}{4} A$. El coeficiente $d(x)$ se escoge a partir de la condición de mínimo de la razón $\eta = \delta/\Delta$, y, por consiguiente, de máximo de $\xi = \gamma_1/\gamma_2$. De resultas se obtiene un algoritmo en el que el número de iteraciones $n_0(\varepsilon)$ depende poco de la razón c_2/c_1. De esto precisamente nos dice la tabla 3.

TABLA 3

$\frac{c_2}{c_1}$	$h = 1/32$		$h = 1/128$	
	$D = E$	$D = d(x)\,E$	$D = E$	$D = d(x)\,E$
2	23	20	45	39
8	46	23	90	47
32	92	25	180	53
128	184	26	360	57
512	367	26	720	59

Capítulo VII

Métodos de diferencias para resolver la ecuación de conductibilidad térmica

En el presente capítulo se examinan los esquemas de diferencias para resolver la ecuación de conductibilidad térmica. Se ha investigado detalladamente una ecuación unidimensional con coeficientes constantes. Se aducen esquemas de diferencias para una ecuación multidimensional de conductibilidad térmica con coeficientes variables.

§ 1. Ecuación de conductibilidad térmica con coeficientes constantes

1. Problema de partida. El proceso de difusión del calor en un vástago unidimensional $0 < x < l$ se describe por la ecuación de conductibilidad térmica

$$c\rho \frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(k \frac{\partial u}{\partial x}\right) + f_0(x, t), \qquad (1)$$

donde $u = u(x, t)$ es la temperatura en el punto x del vástago en el momento t; c es la capacidad calorífica de la unidad de masa, ρ es la densidad de la masa, $c\rho$ es la capacidad calorífica de la unidad de longitud, k es el coeficiente de conductibilidad térmica y f_0, la densidad de las fuentes térmicas. En el caso general k, c, ρ, f_0 pueden depender no sólo de x y t, sino también de la temperatura $u = u(x, t)$ (ecuación casi lineal de conductibilidad térmica) e incluso de $\partial u/\partial x$ (ecuación no lineal). Si k, c, ρ son constantes, entonces (1) puede anotarse en la forma

$$\frac{\partial u}{\partial t} = a^2 \frac{\partial^2 u}{\partial x^2} + f, \qquad f = \frac{f_0}{c\rho}, \qquad (2)$$

donde $a^2 = k/(c\rho)$ es el coeficiente de conducción de temperatura. Sin perjudicar la generalidad de nuestros razona-

mientos podemos considerar $a = 1$, $l = 1$. En efecto, introduciendo las variables $x_1 = \frac{x}{l}$, $t_1 = \frac{a^2 t}{l^2}$, $f_1 = \frac{l^2}{a^2} f$, obtenemos

$$\frac{\partial u}{\partial t_1} = \frac{\partial^2 u}{\partial x_1^2} + f_1, \qquad 0 < x_1 < 1.$$

Se examinará aquí el primer problema de contorno (a veces suele decirse: problema inicial de contorno) en el dominio $\bar{D} = \{0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T\}$. Se pide hallar la solución $u(x, t)$, continua en $\bar{D}$, del problema

$$\begin{gathered} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x, t), \quad 0 < x < 1, \quad 0 < t \leqslant T, \\ u(x, 0) = u_0(x),\ 0 \leqslant x \leqslant 1,\ u(0, t) = u_1(t), \\ u(1, t) = u_2(t), \qquad 0 \leqslant t \leqslant T. \end{gathered} \tag{3}$$

2. Algunas propiedades de las soluciones de la ecuación de conductibilidad térmica. En virtud del principio del máximo, para la solución del problema (3) tiene lugar una estimación

$$\begin{aligned} &\max_{0 \leqslant x \leqslant 1,\ 0 \leqslant t \leqslant T} |u(x, t)| \leqslant \\ &\leqslant \max\left(\max_{0 \leqslant x \leqslant 1} |u_0(x)|,\ \max_{0 \leqslant t \leqslant T} |u_1(t)|,\ \max_{0 \leqslant t \leqslant T} |u_2(t)| \right) + \\ &\qquad\qquad + \int_0^T \max_{0 \leqslant x \leqslant 1} |f(x, t)|\, dt. \end{aligned} \tag{4}$$

Tomemos una ecuación homogénea con las condiciones de contorno homogéneas:

$$\begin{gathered} \frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2}, \qquad 0 < x < 1, \qquad 0 < t < T, \\ u(0, t) = u(1, t) = 0, \qquad 0 \leqslant t \leqslant T, \\ u(x, 0) = u_0(x), \qquad 0 \leqslant x \leqslant 1. \end{gathered} \tag{5}$$

La solución de este problema se halla por el método de separación de variables en la forma

$$u(x, t) = \sum_{k=1}^{\infty} c_k e^{-\lambda_k t} X_k(x). \tag{6}$$

donde λ_k y $X_k(x)$ son los valores propios y las funciones propias ortonormalizadas del problema

$$X'' + \lambda X = 0, \qquad 0 < x < 1, \qquad X(0) = X(1) = 0,$$

iguales a

$$\lambda_k = k^2\pi^2, \qquad X_k(x) = \sqrt{2}\,\text{sen}\, k\pi x, \tag{7}$$

con la particularidad de que

$$(X_k, X_m) = \int_0^1 X_k(x) X_m(x)\, dx = \delta_{km},$$

$$\delta_{km} = \begin{cases} 1, & k = m, \\ 0, & k \neq m. \end{cases}$$

Efectivamente, todas las soluciones particulares (armónicas) $u_k(x, t) = c_k e^{-\lambda_k t} X_k(x)$ satisfacen la ecuación y las condiciones de contorno (5). De la condición inicial

$$u(x, 0) = u_0(x) = \sum_{k=1}^{\infty} c_k X_k(x) \tag{8}$$

se hallan los coeficientes $c_k = (u_0, X_k)$.

De (6) y (8) se infiere

$$\| u(t) \|^2 = (u(x, t), u(x, t)) =$$
$$= \sum_{k=1}^{\infty} c_k^2 e^{-2\lambda_k t} \| X_k \|^2 \leqslant e^{-2\lambda_1 t} \sum_{k=1}^{\infty} c_k^2 = e^{-2\lambda_1 t} \| u_0 \|^2,$$

puesto que

$$\| u_0 \|^2 = \sum_{k=1}^{\infty} c_k^2, \qquad \lambda_k > \lambda_{k-1} > \ldots > \lambda_1 = \pi^2.$$

De este modo, para la solución del problema (5) resulta justa la estimación

$$\| u(t) \| \leqslant e^{-\lambda_1 t} \| u_0 \|, \qquad \lambda_1 = \pi^2, \tag{9}$$

que exprese la propiedad de estabilidad asintótica (para $t \to \infty$) del problema (5) respecto de los datos iniciales (§ 4, p. 7, cap. V). Debido a que $\lambda_k = k^2\pi^2$ crece con el crecimiento de k, a partir de cierto momento t, en la suma (6)

se hará preponderante el primer sumando (primera armónica), es decir, tendrá lugar una igualdad aproximada

$$u(x, t) \approx c_1 e^{-\lambda_1 t} X_1(x).$$

Esta etapa del proceso lleva el nombre de régimen regular.

3. Esquemas de diferencias. En el dominio D introduzcamos una red

$$\bar{\omega}_{h\tau} = \{(x_i t_j): x_i = ih, \quad t_j = j\tau, \quad i = 0, 1, \ldots$$
$$\ldots, N, h = 1/N, \quad j = 0, 1, \ldots, L, \quad \tau = T/L\}$$

con los pasos: h según x y τ según t. Al cambiar la derivada respecto de x por la expresión de diferencias

$$\left(\frac{\partial^2 u}{\partial x^2}\right)_i \sim \frac{u_{i+1} - 2u_i + u_{i-1}}{h^2} = u_{\bar{x}x,\, i} = \Lambda u_i,$$

obtendremos en lugar de (3) un sistema de ecuaciones diferenciales en diferencias (*método de las rectas*)

$$\frac{dv_i}{dt} = \Lambda v_i + f_i, \qquad i = 1, 2, \ldots,$$

con las condiciones de contorno e iniciales

$$v_0(t) = u_1(t), \quad v_N(t) = u_2(t), \quad v_i(0) = u_0(x_i).$$

Para la resolución numérica de este problema sustituyamos, por analogía con el cap. V, la derivada respecto de t por una razón de diferencias

$$\frac{dv_i}{dt} \sim \frac{v_i(t_{j+1}) - v_i(t_j)}{\tau} = \frac{v_i^{j+1} - v_i^j}{\tau} = (v_t)_i^j,$$

y tomemos el segundo miembro en forma de una combinación lineal de valores para $t = t_j$ (en la j-ésima capa) y $t = t_{j+1}$ (en la $(j+1)$-ésima capa):

$$\frac{y_i^{j+1} - y_i^j}{\tau} = \sigma \Lambda y_i^{j+1} + (1 - \sigma) \Lambda y_i^j + \varphi_i^j, \tag{10}$$

donde σ es el parámetro, mientras que φ_i^j es cierto segundo miembro, por ejemplo, $\varphi_i^j = f_i^j$, $\varphi_i^j = f_i^{j+1/2}$, etc. Se deben agregar aquí las condiciones complementarias

$$y_0^j = u_1(t_j), \quad y_N^j = u_2(t_j), \quad y_i^0 = u_0(x_i), \tag{11}$$
$$j = 0, 1, 2, \ldots, \qquad 0 \leqslant i \leqslant N.$$

El esquema (10) está definido en un molde 6-puntual

$$\begin{array}{ccc} (x_{i-1},\ t_{j+1}) & (x_i,\ t_{j+1}) & (x_{i+1},\ t_{j+1}) \\ \times & \times & \times \\ \times & \times & \times \\ (x_{i-1},\ t_j) & (x_i,\ t_j) & (x_{i+1},\ t_j) \end{array}$$

Examinemos un esquema explícito ($\sigma = 0$) en el molde 4-puntual:

$$\frac{y_i^{j+1}-y_i^j}{\tau}=\frac{y_{i-1}^j-2y_i^j+y_{i+1}^j}{h^2}+\varphi_i^j. \qquad (12)$$

Los valores en la $(j+1)$-ésima capa se hallan por la fórmula explícita

$$y_i^{j+1}=\left(1-\frac{2\tau}{h^2}\right)y_i^j+\frac{\tau}{h^2}\left(y_{i-1}^j+y_{i+1}^j\right)+\tau\varphi_i^j.$$

En el caso de $\sigma = 1$ obtenemos un esquema completamente implícito, esto es, un esquema con adelantamiento en el molde $\overset{\bullet\bullet\bullet}{\bullet}$:

$$\frac{y_i^{j+1}-y_i^j}{\tau}=\frac{y_{i-1}^{j+1}-2x_i^{j+1}+y_{i+1}^{j+1}}{h^2}+\varphi_i^j. \qquad (13)$$

Para determinar y_i^{j+1} de (13) obtenemos el problema de contorno

$$\frac{\tau}{h^2}y_{i-1}^{j+1}-\left(1+\frac{2\tau}{h^2}\right)y_i^{j+1}+\frac{\tau}{h^2}y_{i-1}^{j+1}=-F_i^j, \quad 0<i<N,$$

$$F_i^j=y_i^j+\tau\varphi_i^j, \quad y_0^{j+1}=u_i(t_{j+1}), \quad y_N^{j+1}=u_2(t_{j+1}),$$

el cual se resuelve por el método de factorización.

Se usa frecuentemente un esquema implícito simétrico (a veces se denomina *esquema de Crank — Nickolson*) con $\sigma = 1/2$ y un molde $\overset{\bullet\bullet\bullet}{\bullet\bullet\bullet}$:

$$\frac{y_i^{i+1}-y_i^j}{\tau}=\frac{1}{2}\left(\frac{y_{i-1}^{j+1}-2y_i^{j+1}+y_{i+1}^{j+1}}{h^2}+\frac{y_{i-1}^j-2y_i^j+y_{i+1}^j}{h^2}\right)+\varphi_i^j. \qquad (14)$$

Los valores de y^{+1} en la nueva capa se determinan en este caso también por el método de factorización para el problema

de contorno:

$$\frac{\tau}{2h^2} y_{i-1}^{j+1} - \left(1 + \frac{\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{2h^2} y_{i-1}^{j+1} = -F_i^j, \quad 0 < i < N,$$
$$y_0^{j+1} = u_1(t_{j+1}), \qquad y_N^{j+1} = u_2(t_{j+1}), \tag{15}$$
$$F_i^j = \left(1 - \frac{\tau}{h^2}\right) y_i^j + \frac{\tau}{2h^2} (y_{i+1}^j + y_{i+1}^j) + \tau\varphi_i^j.$$

En el caso general (para σ cualquiera) el esquema (10) se denomina *esquema con pesos*. Cuando $\sigma = 0$, el esquema es implícito e y_i^{j+1} se determina por el método de factorización como una solución del problema

$$\sigma\tau\Lambda y_i^{j+1} - y_i^{j+1} = -F_i^j, \qquad 0 < i < N,$$
$$y_0^{j+1} = u_1(t_{j+1}), \quad y_N^{j+1} = u_2(t_{j+1}), \qquad j = 0, 1, \ldots \tag{16}$$

Procedamos ahora al estudio de las propiedades del esquema (10) con σ cualquiera.

4. Estimación del error de aproximación. Con el fin de estimar el orden de exactitud del esquema con pesos (10), se debe estimar primero el error de aproximación (el residuo) y hallar las estimaciones apriorísticas que expresan la estabilidad del esquema respecto del segundo miembro. El esquema de diferencias (10), (11) toma en consideración los datos iniciales y de frontera exactos. Escribamos el esquema (10) en la forma sin índices. Al introducir las designaciones

$$y = y_i^j, \quad \hat{y} = y_i^{j+1}, \quad \Lambda y = y_{\bar{x}x} = \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2},$$
$$y_t = \frac{y_i^{j+1} - y_i^j}{\tau}, \quad y^{(\sigma)} = \sigma y_i^{j+1} + (1 - \sigma) y_i^j,$$

obtenemos

$$y_t = \Lambda y^{(\sigma)} + \varphi, \quad (x_i t_j) \in \omega_{h\tau}, \quad y(x, 0) = u_0(x),$$
$$y_0 = \mu_1(t), \quad y_N = u_2(t) \ (t = t_j = j\tau, \quad j = 0, 1, \ldots). \tag{17}$$

Sea $u = u(x_i, t_j)$ la solución exacta del problema de partida (3) y sea y la solución del problema de diferencias (17).

Sustituyendo en (17) $y = z + u$, obtendremos para el error $z = y - u$ las siguientes condiciones:

$$z_t = \Lambda z^{(\sigma)} + \psi, \qquad (x, t) \in \omega_{h\tau},$$
$$z(x, 0) = 0, \qquad z(0, t) = z(1, t) = 0, \quad (18)$$

donde

$$\psi = \Lambda u^{(\sigma)} + \varphi - u_t \quad (19)$$

es el error de aproximación del esquema (17) en la solución $u = u(x, t)$ del problema (3) (el residuo del esquema).

Encontraremos el desarrollo de ψ según las potencias de h y τ en el entorno del punto $(x_i, \bar{t} = t_j + {}^1/_2\tau)$. Al tomar en consideración que

$$u^{(\sigma)} = \sigma\hat{u} + (1-\sigma)u = \frac{u+\hat{u}}{2} + \left(\sigma - \frac{1}{2}\right)\tau u_t,$$

$$\hat{v} = \bar{v} + \frac{1}{2}\tau\frac{\partial \bar{v}}{\partial t} + \frac{\tau^2}{8}\frac{\partial^2 \bar{v}}{\partial t^2} + \frac{\tau^3}{48}\frac{\partial^3 \bar{v}}{\partial t^3} + O(\tau^4),$$

$$\bar{v} = v\left(x, t_j + \frac{1}{2}\tau\right),$$

$$v = \bar{v} - \frac{1}{2}\tau\frac{\partial \bar{v}}{\partial t} + \frac{\tau^2}{8}\frac{\partial^2 \bar{v}}{\partial t^2} - \frac{\tau^3}{48}\frac{\partial^3 \bar{v}}{\partial t^3} + O(\tau^4),$$

$$\Lambda u = u_{\bar{x}x} = Lu + \frac{h^2}{12}u^{\mathrm{IV}} + O(h^4), \qquad Lu = \frac{\partial^2 u}{\partial x^2},$$

obtenemos

$$\psi = L\bar{u} + \bar{f} - \frac{\partial \bar{u}}{\partial t} + \varphi - \bar{f} + \left(\sigma - \frac{1}{2}\right)\tau L\frac{\partial u}{\partial t} + \frac{h^2}{12}L^2u + O(\tau^2 + h^4).$$

Por cuanto, en virtud de la ecuación (3),

$$L\bar{u} + \bar{f} - \frac{\partial \bar{u}}{\partial t} = 0,$$

entonces

$$L\frac{\partial u}{\partial t} = L^2u + Lf$$

y

$$\psi + \left(\varphi - \bar{f} + \left(\sigma - \frac{1}{2}\right) \tau L \bar{f}\right) + \\ + \left(\frac{h^2}{12} + \left(\sigma - \frac{1}{2}\right) \tau\right) L^2 \bar{u} + O(\tau^2 + h^4).$$

De aquí se ve que

$$\psi = O(\tau + h^2) \text{ para } \varphi = f \text{ y } \sigma \neq \frac{1}{2},$$

$$\psi = O(\tau^2 + h^2) \text{ para } \varphi = \bar{f} \text{ y } \sigma = \frac{1}{2}.$$

Si elegimos σ de un modo tal que el coeficiente de $L^2\bar{u}$ sea nulo:

$$\sigma = \sigma_* = \frac{1}{2} - \frac{h^2}{12\tau}, \tag{20}$$

y ponemos φ igual a

$$\varphi = \bar{f} + \frac{h^2}{12} L\bar{f}, \text{ o bien } \varphi = \bar{f} + \frac{h^2}{12} \Lambda \bar{f} \tag{21}$$

(ambas expresiones se diferencian en una magnitud $O(h^4)$, puesto que $\Lambda f - Lf = O(h^2)$), obtendremos un esquema con el orden de aproximación aumentado (respecto de x): $\psi = O(h^4 + \tau^2)$ para $\sigma = \sigma_*$. Este esquema es también implícito, por lo cual y_i^{j+1} se halla de la ecuación $\sigma_* \tau \Lambda \hat{y} - \hat{y} = -F$ por el método de factorización.

5. Estabilidad del esquema. Volvamos al estudio de la estabilidad y de la convergencia del esquema (17). Veamos, primero, el esquema explícito ($\sigma = 0$) y el esquema implícito puro ($\sigma = 1$). La ecuación (17) para el esquema explícito se escribirá en la forma

$$y_i^{j+1} = \left(1 - \frac{2\tau}{h^2}\right) y_i^j + \frac{\tau}{h^2}(y_{i-1}^j + y_{i+1}^j) + \tau\varphi_i^j, \quad 0 < i < N,$$
$$y_0^{j+1} = 0, \quad y_N^{j+1} = 0, \quad y_i^0 = u_0(x_i), \quad 0 \leqslant i \leqslant N. \tag{22}$$

Si el coeficiente de y_i^j es no negativo, es decir, si

$$\tau \leqslant h^2/2, \tag{23}$$

entonces de (22) proviene que

$$\| y^{j+1} \|_C \leqslant \| y^j \|_C + \tau \| \varphi^j \|_C, \tag{24}$$

donde $\| y \|_C = \max\limits_{0 \leqslant i \leqslant N} | y_i |$. La sumación según k de 0 a $j - 1$ nos da

$$\| y^j \|_C \leqslant \| y^0 \|_C + \sum_{k=0}^{j-1} \tau \| \varphi^k \|_C. \tag{25}$$

Esta desigualdad expresa precisamente la estabilidad en una norma reticular C del esquema explícito respecto de los datos iniciales y del segundo miembro bajo la condición (23) (el esquema explícito es condicionalmente estable).

El esquema implícito (17) con $\sigma = 1$ se escribirá en la forma

$$\tau \Lambda y_i^{j+1} - y_i^{j+1} = -F_i, \qquad F_i = y_i^j + \tau \varphi_i^j$$

o bien

$$\frac{\tau}{h^2} y_{i-1}^{j+1} - \left(1 + \frac{2\tau}{h^2}\right) y_i^{j+1} + \frac{\tau}{h^2} y_{i+1}^{j+1} = -F_i^j, \quad 0 < i < N,$$
$$y_0^{j+1} = y_N^{j+1} = 0.$$

Ahora, hagamos uso del teorema 3 del § 5, cap. I: para la solución del problema

$$A_i y_{i-1} - C_i y_i + A_{i+1} y_{i+1} = -F_i,$$
$$C_i = A_i + A_{i+1} + D_i, \quad 0 < i < N, \quad y_0 = y_N = 0$$

es justa la estimación

$$\| y \|_C \leqslant \left\| \frac{F}{D} \right\|_C.$$

En nuestro caso $A_i = A_{i+1} = \tau/h^2$, $D_i = 1$,

$$\| y^{k+1} \|_C \leqslant \| F^k \|_C \leqslant \| y^k \|_C + \tau \| \varphi^k \|_C. \tag{26}$$

De aquí, sumando según $k = 0, 1, \ldots, j - 1$, obtenemos la estimación (25). De este modo, un esquema implícito puro es incondicionalmente estable, es decir, es estable para cualesquiera τ y h. Siendo σ arbitrario, la ecuación en di-

ferencias tiene por expresión

$$\frac{\sigma\tau}{h^2} y_{i-1}^{j+1} - \left(1 + \frac{2\sigma\tau}{h^2}\right) y_i^{j+1} + \frac{\sigma\tau}{h^2} y_{i+1}^{j+1} = -F_{i x}^j$$

$$0 < i < N, \qquad y_0^{j+1} = y_N^{j+1} = 0,$$

$$F_i^j = \left(1 - \frac{2(1-\sigma)\tau}{h^2}\right) y_i^j + \frac{(1-\sigma)}{h^2} \tau (y_{i-1}^j + y_{i+1}^j) + \tau\varphi_i^j.$$

De aquí se ve que el coeficiente de y_i^j es no negativo, si

$$\tau \leqslant \frac{h^2}{2(1-\sigma)} \text{ o bien } \sigma \geqslant 1 - \frac{h^2}{2\tau}. \tag{27}$$

Bajo esta condición $\| F \|_C \leqslant \| y \|_C + \tau \| \varphi \|_C$; aprovechando, luego, el teorema 3 del § 5, cap. I, obtendremos la estimación (25) para la condición (27). En particular, para un esquema simétrico la estabilidad en C tiene lugar cuando $\tau \leqslant h^2$. En realidad el esquema (17), con $\sigma \geqslant 1/2$ es incondicionalmente estable en C respecto de los datos iniciales, de suerte que

$$\| y^j \|_C \leqslant M_0 \| \dot{y} \|_C,$$

donde $M_0 = \text{const} > 1$. No obstante, esta desigualdad se demuestra de un modo bastante complejo.

Más abajo se probará que en otra norma la condición de estabilidad de un esquema con pesos tiene por expresión

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{h^2}{4\tau}, \tag{28}$$

de modo que el esquema con $\sigma \geqslant 1/2$ es incondicionalmente estable, mientras que para $\sigma < 1/2$ la condición de estabilidad, en lugar de (27), se expresa así

$$\tau \leqslant \frac{h^2}{4(1/2-\sigma)}. \tag{29}$$

El resultado obtenido (29) se obtiene a base de la teoría general de estabilidad.

Por analogía con el § 4, cap. I, introduzcamos un operador A:

$$Ay = -\Lambda \dot{y}, \qquad y \in \Omega, \qquad \dot{y} \in \dot{\Omega},$$

donde $\dot{\Omega}$ es un conjunto de funciones y definidas sobre la red $\dot{\omega}_h = \{x_i: x_i = ih, i = 0, 1, \ldots, N, h = 1/N\}$ e igua-

les a cero en la frontera para $i = 0, N$; o y es un conjunto de funciones definidas en los nodos interiores de la red $x \in \in \omega_h = \{x_i : x_i = ih, i = 1, 2, \ldots, N-1, h = 1/N\}$.

Escribamos el esquema con pesos en la forma canónica:

$$Bz_t + Az = \psi(t), \ t \in \bar{\omega}_\tau, \quad Z(0) = 0, \ B = E + \sigma\tau A. \tag{30}$$

Con este fin resulta suficiente sustituir $z^{(\sigma)} = \sigma\hat{z} + (1 - \sigma) z = z + \sigma(\hat{z} - z) = z + \sigma\tau z_t$ en (18).

El operador A es, de acuerdo con lo mostrado en el cap. I, autoconjugado y positivo: $A = A^* > 0$, si el producto escalar en H lo definimos según la fórmula

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h.$$

La estabilidad del esquema (30) fue investigada en el cap. V, donde probamos que el esquema (30) es estable en H_A cuando

$$\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau \| A \|}. \tag{31}$$

En nuestro caso $\| A \| = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}$. De aquí proviene que el esquema (17) es estable para cualesquiera τ y h, siempre que $\sigma \geqslant 1/2$. Si $\sigma < 1/2$, el esquema será estable para

$$\tau \leqslant \frac{1}{(1/2 - \sigma) \| A \|}.$$

Al sustituir aquí $\| A \| \approx 4/h^2$, obtenemos

$$\tau \leqslant \frac{h^2}{4(1/2 - \sigma)} \quad \text{y} \quad 4(1/2 - \sigma)\tau \leqslant h^2.$$

En particular, cuando $\sigma = \sigma_*$, tenemos $4(1/2 - \sigma_*)\tau = h^2/3 < h^2$, es decir, el esquema con un orden de aproximación aumentado es incondicionalmente estable.

6. Convergencia del esquema. Con el fin de demostrar la convergencia del esquema (17) se debe obtener la estimación apriorística para el problema (30). Hagamos uso de la desigualdad para z, obtenida al investigar la convergencia de los esquemas en el cap. V, en virtud de la cual para (30)

y (18) es justa la estimación del error

$$\| z^j \|_A \leqslant \sum_{k=0}^{j-1} \tau \| \psi^k \| \quad \text{para } \sigma \geqslant 0,\ \sigma \geqslant \sigma_0. \tag{32}$$

Al sustituir aquí $Az = \overset{\circ}{z}_{\bar{x}x}$, hallaremos

$$\| z \|_A^2 = (Az, z) = -(\overset{\circ}{z}_{\bar{x}, x}, \overset{\circ}{z}) = (\overset{\circ}{z}_{\bar{x}}, \overset{\circ}{z}_{\bar{x}}] = \sum_{i=1}^{N} h (z_{\bar{x}, i})^2$$

y aprovecharemos la estimación

$$\| z \|_C = \max_{x \in \omega_n} | z | \leqslant \frac{1}{2} \left(\sum_{i=1}^{N} h (z_{\bar{x}, i})^2 \right)^{1/2} = \frac{1}{2} \| z \|_A.$$

De resultas obtenemos

$$\| z^j \|_C \leqslant \frac{1}{2} \sum_{k=0}^{j-1} \tau \| \psi^k \|, \tag{33}$$

es decir, el esquema (17) converge en la norma reticular C con la velocidad $\| y^j - u^j \|_C = \| z^j \|_C = O(h^2 + \tau)$ para $\sigma \neq 1/2$, $\sigma \geqslant \sigma_0$, $\| z^j \|_C = O(h^2 + \tau^2)$, $\sigma = 1/2$. Si $\sigma_* \geqslant 0$, es decir, si $\tau \geqslant h^2/\sigma$, entonces para el esquema $\sigma = \sigma_*$ es también justa la estimación (33) y

$$\| z^j \|_C = O(h^4 + \tau^2) \quad \text{para } \sigma = \sigma_*.$$

7. Estabilidad asintótica. La propiedad de estabilidad *asintótica (para $t \to \infty$) del problema* (5) respecto de los datos iniciales se expresa por la estimación (9). Para t grandes la solución del problema (5) se determina por la primera armónica

$$u(x, t) \approx c_1 e^{-\lambda_1 t} X_1(x)$$

(*régimen regular*). *Es natural exigir que la solución del problema de diferencias*

$$\begin{gathered} y = \sigma \Lambda \hat{y} + (1 - \sigma) \Lambda y; \quad x = ih, \quad t = j\tau, \\ i = 1, 2, \ldots, N-1, \quad j = 0, 1, \ldots, \\ y(0, t) = 0,\ y(1, t) = 0, \quad y(x, 0) = u_0(x) \end{gathered} \tag{34}$$

posea las propiedades analíticas.

En el cap. V para el esquema operacional de diferencias con pesos

$$By_t + Ay = 0, \quad t \in \omega_\tau, \quad y(0) = y_0, \quad B = E + \sigma\tau A,$$
$$\delta E \leqslant A \leqslant \Delta E, \quad \delta > 0, \quad A = A^* > 0$$

se ha establecido la estabilidad asintótica del esquema con pesos

$$\|y^j\| \leqslant e^{-\delta t_j} \|\overset{\circ}{y}\|$$

con la condición complementaria

$$\tau \leqslant \tau_0(\sigma)$$

donde $\tau_0 = 2/(\delta + \Delta)$ para un esquema explícito ($\sigma = 0$), $\tau_0 = \infty$ (τ es cualquiera) para un esquema implícito ($\sigma = 1$) y $\tau_0 = 2/\sqrt{\delta\Delta}$ para un esquema simétrico ($\sigma = 1/2$). Para el esquema (34) tendremos

$$\delta = \frac{4}{h^2} \operatorname{sen}^2 \frac{\pi h}{2}, \quad \Delta = \frac{4}{h^2} \cos^2 \frac{\pi h}{2}, \quad \delta + \Delta = \frac{4}{h^2}.$$

Para un esquema explícito ($\sigma = 0$) $\tau_0 = h^2/2$ y la condición de estabilidad asintótica coincide con la de estabilidad corriente; el esquema implícito ($\sigma = 1$) es, como antes, incondicionalmente estable. Sin embargo, el esquema simétrico ($\sigma = 1/2$) será incondicionalmente estable en el sentido habitual y asintóticamente estable bajo la condición

$$\tau \leqslant \tau_0, \quad \tau_0 = \frac{h^2}{\operatorname{tg} \pi h} \approx \frac{h}{\pi}.$$

En este caso la solución del problema de diferencias (34) con $\sigma = 1/2$, para t grandes, se determina por la primera armónica:

$$y_i^j \approx c_1 \rho^j \operatorname{sen} \pi x_i \approx c_1 e^{-\lambda_1 t_j} \operatorname{sen} \pi x.$$

Aquí $\rho = (1 - 1/2\tau\delta)/(1 + 1/2\tau\delta) = e^{-\lambda_1 t}(1 + O(\tau^2))$.

Si la condición $\tau \leqslant \tau_0$ está perturbada, es decir, si $\tau > \tau_0$, entonces para t grandes predomina no la primera, sino la última armónica:

$$y_i^j \approx c_1 \rho^j \operatorname{sen} \pi (N - 1) x_i \approx c_1 \rho^j (-1)^i \operatorname{sen} \pi x_i,$$

donde $\rho = \dfrac{1/2\tau\Delta - 1}{1/2\tau\Delta + 1} < e^{-\lambda_1 \tau}$, lo que, por supuesto, no tiene nada en común con la solución de una ecuación diferencial.

La exigencia de la estabilidad asintótica está estrechamente ligada con la exactitud del esquema y de hecho significa también la exigencia de exactitud asintótica. Esto se pone de manifiesto con mayor claridad en los cálculos sobre las redes reales para t grandes. Hemos de observar que la condición $\tau \approx h/\pi$ para un esquema simétrico no parece abrumadora. Se demuestra que un esquema implícito puro ($\sigma = 1$) puede asegurar una exactitud admisible para grandes valores de t sólo cuando el paso τ sea comparable con el de un esquema explícito, con lo que en los cálculos para t grandes el esquema implícito puro queda privado de su ventaja principal, a saber, la estabilidad para τ y h cualesquiera.

§ 2. Problemas multidimensionales de la conductibilidad térmica

1. Esquema de diferencias con pesos. Examinemos en un plano $x = (x_1, x_2)$ un dominio G con la frontera Γ. Buscaremos la solución del problema de conductibilidad térmica en el dominio $\bar{G} = G + \Gamma$ para todo $0 \leqslant t \leqslant T$. Se pide hallar una función $u(x, t)$ que esté definida en el cilindro $\overline{Q_T} = \bar{G} \times [0, T] = \{(x, t): x \in G, 0 \leqslant t \leqslant T\}$ y que satisfaga en $Q_T = G \times (0, T] = \{(x, t): x \in G, 0 < t \leqslant T\}$ la ecuación de conductibilidad térmica

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \quad Lu = \frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}, \tag{1}$$

las condiciones de contorno de primera especie en la frontera Γ del dominio G

$$u = \mu(x, t), \quad x \in \Gamma, \quad 0 \leqslant t \leqslant T, \tag{2}$$

y la condición inicial para $t = 0$:

$$u(x, 0) = u_0(x), \quad x \in \bar{G}. \tag{3}$$

Supongamos que $\bar{G}$ es un rectángulo:

$$\bar{G} = \{x = (x_1, x_2): 0 \leqslant x_1 \leqslant l_1, \quad 0 \leqslant x_2 \leqslant l_2\}.$$

Introduzcamos en $\bar{G}$ una red rectangular

$$\bar{\omega}_h = \{x_i = (x_1^{(i_1)}, x_2^{(i_2)}): x_\alpha^{(i_\alpha)} = i_\alpha h_\alpha,$$
$$i_\alpha = 0, 1, \ldots, N_\alpha, \ h_\alpha = l_\alpha/N_\alpha, \ \alpha = 1, 2\}$$

con la frontera

$$\gamma_h = \{x_i = (i_1h_1,\ i_2h_2):\ i_1 = 0,\ N_1,\ 0 < i_2 < N_2;\ i_2 = 0,\ N_2,\ 0 < i_1 < N_1\}.$$

Aproximemos el operador de Laplace $Lu = \Delta u$ mediante un operador de diferencias en el molde pentapuntual (véase el cap. VI, § 1)

$$Lu \sim \Lambda u = u_{\bar{x}_1 x_1} + u_{\bar{x}_2} u_{x_2}.$$

Sustituyamos el problema (1) — (3) por un problema diferencial en diferencias (método de las rectas):

$$\frac{dv_i(t)}{dt} = \Lambda v_i(t) + f_i(t),\ i = (i_1,\ i_2),\ v_i(0) = u_0(x_i),$$
$$x_i \in \omega_h,\quad v_i(t)|_{\gamma_h} = \mu_i(t),\quad 0 \leqslant t \leqslant . \tag{4}$$

Introduzcamos en el segmento $0 \leqslant t \leqslant T$ una red $\bar{\omega}_\tau = \{t_j = j\tau:\ 0 \leqslant t_j \leqslant T\}$ de paso τ. Escribamos un esquema con pesos

$$\frac{y^{j+1} - y^j}{\tau} = \Lambda\ (\sigma y^{j+1}) + (1-\sigma)\ y^j) + \varphi^j,\quad j = 0,\ 1,\ \ldots, \tag{5}$$

donde $y^j = y(x_i,\ t_j) = y(i_1h_1,\ i_2h_2;\ t_j)$, $x = (i_1h_1,\ i_2h_2) \in \in \omega_h$. Agregamos a las ecuaciones (5)

$$y(x,\ 0) = u_0(x),\quad x = (i_1h_1,\ i_2h_2) \in \bar{\omega}_h,$$
$$y(x_1,\ t) = \mu_i(t),\quad x \in \gamma_h,\ t = j\tau \in \omega_h. \tag{5'}$$

De aquí se ve que para determinar $\hat{y} = y^{j+1}$ en una capa nueva $t = t_{j+1}$ se debe resolver una ecuación en diferencias

$$\hat{y} - \sigma\tau\Lambda\hat{y} = F,\ F = y + (1 - \sigma)\ \tau\Lambda y + \tau\varphi,\ x \in \omega_h,$$
$$\hat{y} = \mu,\quad x \in \gamma_h. \tag{6}$$

La resolubilidad de este problema se desprende de lo que el operador $(E - \sigma\tau\Lambda)$ es definido positivo para $\sigma > -1/(\tau\ \|A\|)$, puesto que $(E - \sigma\tau\Lambda)\ \mathring{y} = (E + \sigma\tau A)\ y$ en el espacio de funciones reticulares $\mathring{y}$ que vienen dadas en la red $\bar{\omega}_h$ y que se reducen a cero en la frontera γ_h (compárese con el cap. VI). Mostrémoslo.

Introduciendo un producto escalar

$$(y, v) = \sum_{x_j \in \omega_h} y(x_i)\, v(x_i)\, h_1 h_2 = \\ = \sum_{i_1=1}^{N_1-1} h_1 \sum_{i_2=1}^{N_2-1} h_2 y(i_1h_1, i_2h_2)\, v(i_1h_1, i_2h_2) \qquad (7)$$

y teniendo presente que $(Ay, y) \leqslant \| Ay \| \; \| y \| \leqslant \| A \| \cdot \cdot \| y \|^2$, encontramos

$$((E - \sigma\tau\Lambda)\, \overset{\circ}{y}, \overset{\circ}{y}) = ((E + \sigma\tau A)\, y, y) = \\ = \|y\|^2 + \sigma\tau\,(Ay, y) \geqslant \left(\frac{1}{\|A\|} + \sigma\tau\right)(Ay, y) > 0,$$

puesto que $(Ay, y) \geqslant \delta \, \| y \|^2 > 0$ (véase el cáp. VI, § 1, p. 5).

Escribamos detalladamente en la forma de índices una ecuación en diferencias

$$\sigma\gamma_1 (\hat{y}_{i_1-1,\, i_2} + \hat{y}_{i_1+1,\, i_2}) - (1 + 2\sigma(\gamma_1 + \gamma_2)) \times \\ \times \hat{y}_{i_1 i_2} + \sigma\gamma_2 (\hat{y}_{i_1 i_2-1} + \hat{y}_{i_1 i_2+1}) = -F_{i_1 i_2}, \qquad (8)$$

donde

$$y_{i_1 i_2} = y(i_1h_1, i_2h_2), \quad \gamma_1 = \tau/h_1^2, \quad \gamma_2 = \tau/h_2^2,$$

$$F_{i_1 i_2} = (1 - 2(1-\tau)(\gamma_1 + \gamma_2))\, y_{i_1 i_2} + (1-\sigma) \times \\ \times \gamma_1 (y_{i_1-1,\, i_2} + y_{i_1+1,\, i_2}) + (1-\sigma)\,\gamma_2 \times \\ \times (y_{i_1,\, i_2-1} + y_{i_1 i_2+1}) + \varphi_{i_1 i_2},$$

$$\hat{y}_{i_1 i_2} = \hat{\mu}_{i_1 i_2}, x_i = (i_1h_1, i_2h_2) \in \gamma_h.$$

Dicho problema de contorno en diferencias se resuelve respecto de $\hat{y}$ por los mismos métodos que se usan en la resolución del problema de Dirichlet en diferencias para la ecuación de Poisson (véase el cap. VI, § 2). Aquí los coeficientes de la ecuación son constantes, el dominio G es un rectángulo, razón por la cual los métodos directos de resolución de las ecuaciones en diferencias (8) resultan los más económicos. Los métodos iterativos son menos económicos.

2. Estabilidad y convergencia. Haciendo uso del operador A, definido anteriormente en el cap. VI:

$$Ay = -\Lambda\overset{\circ}{y} = -\overset{\circ}{y}_{\bar{x}_1x_1} - \overset{\circ}{y}_{\bar{x}_2x_2}, \quad \overset{\circ}{y} \in \overset{\circ}{\Omega}, \quad y \in \Omega = H,$$

escribamos el esquema (5) en la forma canónica:

$$B\frac{y^{j+1}-y^j}{\tau} + Ay^j = \varphi^j, \quad j=0, 1, \ldots, \overset{\circ}{y} = u_0, \quad y \in H,$$

$$B = E + \sigma\tau A. \tag{9}$$

El operador A fue estudiado en el cap. VI. Es autoconjugado y definido positivo en el espacio $H = \Omega$ de dimensión

$$(N_1 - 1)(N_2 - 1), \; A = A^*, \; \delta_0 E \leqslant A \leqslant \Delta_0 E,$$

donde

$$\delta_0 = \frac{4}{h_1^2}\operatorname{sen}^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\operatorname{sen}^2\frac{\pi h_2}{2l_2};$$

$$\Delta_0 = \frac{4}{h_1^2}\cos^2\frac{\pi h_1}{2l_1} + \frac{4}{h_2^2}\cos^2\frac{\pi h_2}{2l_2}, \quad \Delta_0 = \|A\|. \tag{10}$$

En virtud de la teoría general (véase el cap. V), el esquema (9) es estable en H_A cuando

$$\delta \geqslant \delta_0, \qquad \delta_0 = \frac{1}{2} - \frac{1}{\tau\|A\|}. \tag{11}$$

En particular, para un esquema explícito tenemos la condición

$$\tau \leqslant \frac{2}{\Delta_0}, \quad \text{o bien } \tau < \left(\frac{2}{h_1^2} + \frac{2}{h_2^2}\right)^{-1}. \tag{12}$$

En la red cuadrada ($h_1 = h_2 = h$) la condición de estabilidad del esquema explícito tiene por expresión

$$\tau < h^2/4$$

(compárese con las condiciones $\tau < h^2/2$ para el problema unidimensional). De (11) se ve que los esquemas con

$$\sigma \geqslant 1/2,$$

incluidos el esquema implícito ($\sigma = 1$) y el simétrico ($\sigma = 1/2$), son incondicionalmente estables. El esquema ex-

plícito ($\sigma = 0$) puede escribirse en la forma

$$y_{i_1 i_2}^{j+1} = (1 - 2(\gamma_1 + \gamma_2))\, y_{i_1 i_2}^{j} + \gamma_1 (y_{i_1-1,\, i_2}^{j} + y_{i_1+1,\, i_2}^{j}) + \\ + \gamma_2 (y_{i_1,\, i_2-1}^{j} + y_{i_1,\, i_2+1}^{j}) + \tau \varphi_{i_1 i_2}^{j}. \qquad (13)$$

La suma de los coeficientes de y en el segundo miembro de (12) es igual a uno. Si todos los coeficientes son no negativos, es decir, si se cumple la condición $\gamma_1 + \gamma_2 \leqslant 1/2$, $\gamma_1 = \tau/h_1^2$, $\gamma_2 = \tau/h_2^2$, equivalente a la condición de estabilidad (12), entonces de (13) proviene una desigualdad

$$\| y^{j+1} \|_C \leqslant \| y^j \|_C + \tau \, \| \varphi^j \|_C.$$

Al sumar según $k = 0, 1, \ldots, j-1$, obtenemos la estimación (compárese con el § 1)

$$\| y^j \|_C \leqslant \| y^0 \|_C + \sum_{k=0}^{j-1} \tau \| \varphi^k \|_C, \qquad (14)$$

que queda en vigor para cualesquiera pasos de la red si el esquema es implícito puro ($\sigma = 1$). En todos los demás casos la estimación (14) tiene lugar para $\sigma \geqslant 1 - 1/\tau\Delta_0$. Para demostrar la convergencia se debe, como siempre, investigar el residuo

$$\psi = \Lambda (\sigma \hat{u} + (1 - \sigma)\, u) + \varphi - u_t.$$

Teniendo presente que $\Lambda u = Lu + O(|h|^2)$, $|h|^2 = h_1^2 + h_2^2$, encontramos, por analogía con el caso unidimensional,

$$\psi = O(|h|^2 + \tau^2) + \left(\tau - \frac{1}{2}\right) O(\tau).$$

Para el error $z = y - u$ tenemos un problema

$$B \frac{z^{j+1} - z^j}{\tau} + Az^j = \psi^j, \quad j = 0, 1, \ldots, \quad z^0 = z(0) = 0.$$

De aquí y de las estimaciones apriorísticas proviene la convergencia en C del esquema (5) con la velocidad $O(\tau + |h|^2)$ para $\sigma \neq 1/2$, y $O(\tau^2 + |h|^2)$ para $\sigma = 1/2$ (analogía completa con el caso unidimensional), siempre que $\sigma \geqslant 1 - \frac{1}{\tau\Delta_0}$.

Para la solución del problema se cumple, en virtud de la estimación obtenida en el cap. V, una desigualdad

$$\|z^{j+1}\|_A \leqslant \sum_{k=0}^{j} \tau\|\psi^k\| \text{ para } \sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau\Delta_0}, \quad \sigma \geqslant 0,$$

donde

$$\|z\|_A^2 = \|z\|_{A_1}^2 + \|z\|_{A_2}^2, \quad A_1 y = -y_{\bar{x}_1 x_1}, \quad A_2 y = -y_{\bar{x}_2 x_2},$$

$$\|z\|_A^2 = \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2-1} h_2 (z_{\bar{x}_1}(i_1, i_2))^2 + \sum_{i_1=1}^{N_1-1} \times$$
$$\times \sum_{i_2=1}^{N_2} h_2 (z_{\bar{x}_2}(i_1, i_2))^2.$$

De aquí proviene la estabilidad incondicional de la convergencia del esquema (5) en H_A con la velocidad $O(\tau + |h|^2)$ para $\sigma \neq 1/2$, $\sigma \geqslant 1/2$, y $O(\tau^2 + |h|^2)$ para $\sigma = 1/2$.

La investigación realizada más arriba debe ser complementada con las condiciones de estabilidad asintótica. Por cuanto las condiciones citadas $\tau \leqslant \tau_0$ se han obtenido para un esquema operacional en diferencias con pesos y operador arbitrario

$$A = A^* > 0, \qquad \delta_0 E \leqslant A \leqslant \Delta_0 E,$$

pueden ser aplicadas, pues, también para nuestro esquema (5). Haciendo uso de las expresiones (10) para δ_0 y Δ_0, obtenemos las condiciones de estabilidad asintótica $\tau \leqslant \tau_0^{(1)}$, $\tau_0^{(1)} = 2\left(\frac{4}{h_1^2} + \frac{4}{h_2^2}\right)^{-1}$ para un esquema explícito ($\sigma = 0$) $\tau \leqslant \tau_0^{(2)}$, $\tau_0^{(2)} = \frac{2}{\sqrt{\delta_0\Delta_0}}$, δ_0, Δ_0 de (10), para un esquema simétrico ($\sigma = 1/2$).

En particular, para $h_1 = h_2 = h$, $l_1 = l_2 = l$ tenemos

$$\delta_0 = \frac{8}{h^2} \operatorname{sen}^2 \frac{\pi h}{2l}, \quad \Delta_0 = \frac{8}{h^2} \cos^2 \frac{\pi h}{2l}, \quad \tau_0^{(1)} = \frac{h^2}{4},$$
$$\tau_0^{(2)} = \frac{h^2}{2}\left(\operatorname{sen} \frac{\pi h}{l}\right)^{-1} \approx \frac{hl}{2\pi}.$$

El valor límite de $\tau_0^{(2)}$ es dos veces menor que para el esquema unidimensional (5) del § 1.

Un esquema implícito puro ($\sigma = 1$) es absolutamente estable asintóticamente.

3. Coeficientes variables. Examinemos el problema (1) suponiendo que L es un operador elíptico de segundo orden, con coeficientes variables, privado de derivadas mixtas:

$$Lu = L_1u + L_2u,\quad L_1u = \frac{\partial}{\partial x_1}\left(k_1(x,\ t)\frac{\partial u}{\partial x_1}\right),$$

$$L_2u = \frac{\partial}{\partial x_2}\left(k_2(x,\ t)\frac{\partial u}{\partial x_2}\right),$$

$$c_1 \leqslant k_\alpha(x,\ t) \leqslant c_2,\qquad (x,\ t) \in \bar{Q}_T = G \times (0,\ T).$$

Aproximemos cada uno de los operadores L_1 y L_2 mediante un operador de diferencias tripuntual:

$$L_1 \sim \Lambda_1,\qquad L_2 \sim \Lambda_2,$$

$$\Lambda_1 v = (a_1 v_{\bar{x}_1})_{x_1},\qquad \Lambda_2 v = (a_2 v_{\bar{x}_2})\,x_2,$$

donde $a_1 = a_1(i_1h_1,\ i_2h_2,\ t)$, $a_2 = a_2(i_1h_1,\ i_2h_2,\ t)$ son ciertas funcionales de los valores k_1 y k_2, respectivamente; en un caso más simple $a_1 = k_1((i_1 - 1/2)\,h_1,\ i_2h_2,\ t)$, $a_2 = k_2(i_1h_1;\ (i_2 - 1/2)\,h_2,\ t)$, lo que asegura el segundo orden de aproximación: $\Lambda_\alpha u - L_\alpha u = O(h_\alpha^2)$, $\alpha = 1,\ 2$. Al operador L se le pone en correspondencia el operador de diferencias Λ:

$$\Lambda v = \Lambda_1 v + \Lambda_2 v = (a_1 v_{\bar{x}_1})_{x_1} + (a_2 v_{\bar{x}_2})_{x_2}. \tag{15}$$

Escribamos $\Lambda_1 v$ y $\Lambda_2 v$ en forma de índices

$$\Lambda_1 v = \frac{1}{h_1}\Big[a_1((i_1+1)\,h_1,\ i_2h_2;\ t)\frac{v_{i_1+1,\ i_2} - v_{i_1i_2}}{h_1} - a_1(i_1h_1,\ i_2h_2;\ t)\frac{v_{i_1i_2} - v_{i_1-1,\ i_2}}{h_1}\Big],$$

$$\Lambda_2 v = \frac{1}{h_2}\Big[a_2(i_1h_1,\ (i_2+1)\,h_2;\ t)\frac{v_{i_2+1} - v_{i_1i_2}}{h_2} - a_2(i_1h_1,\ i_2h_2;\ t)\frac{v_{i_1i_2} - v_{i_1,\ i_2-1}}{h_2}\Big].$$

El esquema de diferencias con pesos tiene la misma expresión (5) que en el p. 1. Se escoge el mismo espacio reticular $H = \Omega$ con producto escalar (7) y se introduce el operador A:

$$Ay = -\Lambda \overset{\circ}{y} = -(a_1 \overset{\circ}{y}_{\bar{x}_1})_{x_1} - (a_2 \overset{\circ}{y}_{\bar{x}_2})_{x_2}.$$

Teniendo en cuenta que para el caso unidimensional del operador A: $Ay = -(a\overset{\circ}{y}_{\bar{x}})_x$

$$c_1 (\overset{\circ}{A} y, y) \leqslant (Ay, y) \leqslant c_2 (\overset{\circ}{A} y, y), \quad \overset{\circ}{A} y = -\overset{\circ}{y}_{\bar{x}x},$$
$$0 < c_1 \leqslant a \leqslant c_2,$$

no es difícil convencerse de que las desigualdades semejantes se verifican también para el operador bidimensional (15):

$$c_1 \overset{\circ}{A} \leqslant A \leqslant c_2 \overset{\circ}{A}, \quad \overset{\circ}{A} y = -\overset{\circ}{y}_{\bar{x}_1 x_1} - \overset{\circ}{y}_{\bar{x}_2 x_2}.$$

De aquí se ve que $\delta E \leqslant A \leqslant \Delta E$, $\delta = c_1 \delta_0$, $\Delta = c_2 \Delta_0$, donde δ_0 y Δ_0 se determinan según las fórmulas (10). Para hallar $\hat{y} = y^{j+1}$ sobre la capa nueva obtenemos el problema (6), en el que Λ se determina de (15). En el caso de un esquema explícito $\hat{y}$ se determina en cada nodo $x \in \omega_h$ por la fórmula

$$\hat{y} = y + (1 - \sigma)\, \tau \Lambda y + \tau \varphi.$$

Para los esquemas implícitos ($\sigma \neq 0$) se debe resolver una ecuación en diferencias pentapuntual con coeficientes variables. Aquí se emplean los métodos iterativos, de los cuales el método alternado triangular resulta ser más económico (véase el cap. V, § 5); el número de iteraciones para dicho método se expresa por la magnitud $O\left(\frac{1}{\sqrt{h}} \ln \frac{1}{\varepsilon}\right)$, siempre que ($\tau \approx O(h)$. La descripción del método alternado triangular para las ecuaciones en diferencias con coeficientes variables se ha dado en el cap. VI; si se aplica a la ecuación (6) con el operador Λ del tipo (15), debe ser un tanto modificado.

§ 3. Esquemas económicos

1. Método de direcciones variables. Al comparar los esquemas explícitos e implícitos (5), detendrémonos en dos características: el volumen de los cálculos para hallar y^{j+1} y las restricciones que se imponen sobre el paso τ.

ESQUEMA EXPLÍCITO: para determinar y^{j+1} sobre la red ω_h hay que realizar un número de operaciones proporcional al número de nodos, es decir, el número de operaciones correspondientes a un nodo no depende de la red ω_h. Sin embargo, el paso τ está rígidamente limitado superiormente por la condición $\tau \leqslant \tau_0(h)$: $\tau \leqslant h^2/4$ con $h_1 = h_2 = h$ para el esquema (13).

ESQUEMA IMPLÍCITO ($\sigma \geqslant 1/2$); para determinar y^{j+1} se debe resolver un sistema de $(N_1 - 1)(N_2 - 1)$ ecuaciones en diferencias pentapuntuales; con este objeto, por lo menos en el caso de coeficientes variables, se requiere un número de operaciones (correspondientes a un nodo de la red ω_h) que crece cuando $|h| \to 0$.

Surge un problema: construir los esquemas en que se combinen las mejores cualidades de esquemas explícitos e implícitos, es decir, los esquemas a construir deben ser incondicionalmente estables con un número de operaciones en cada capa proporcional al número de nodos de la red ω_h. Los esquemas de este género suelen llamarse *económicos*. Por supuesto, hemos de hacer una especificación de que los esquemas incondicionalmente estables en el sentido habitual han de ser asintóticamente estables, lo que conduce a una restricción del paso mucho más débil (por ejemplo, $\tau \leqslant lh/(2\pi)$ para $\sigma = 1/2$, $h_1 = h_2 = h$, $l_1 = l_2 = l$) que la condición de estabilidad ($\tau \leqslant h^2/4$) para el esquema explícito. Además, la condición $\tau = O(h)$ es natural para el esquema $O(\tau^2 + |h|^2)$.

Los primeros esquemas económicos han aparecido en los años 1955—1956 y se han denominado *métodos de direcciones variables*. La idea algorítmica principal (en la que se radica la economía) consiste en que para pasar de una capa t_j a otra capa t_{j+1} se deben resolver, por el método de factorización, las ecuaciones en diferencias tripuntuales, primero a lo largo de las filas y, a continuación, a lo largo de las columnas de la red ω_h.

Demos a conocer las fórmulas del método de direcciones variables (*esquema* longitudinal transversal *de Pismann-Recford*) para el problema (1) con un operador L: $Lu = L_1 u + L_2 u$, donde L_α es uno de los operadores:

$$L_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2} \text{ o bien } L_\alpha u = \frac{\partial}{\partial x_\alpha}\left(k_\alpha (x, t) \frac{\partial u}{\partial x_\alpha}\right), \quad \alpha = 1, 2.$$

Sean Λ_1, Λ_2 los operadores tripuntuales correspondientes y supongamos que $\Lambda = \Lambda_1 + \Lambda_2$. Introduciendo un valor intermedio de $\bar{y} = y^{j+1/2}$, enunciamos el esquema de diferencias de direcciones variables:

$$\frac{y^{j+1/2} - y^j}{\tau/2} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^j + \varphi^j, \quad x \in \omega_h,$$

$$y^{j+1/2} = \bar{\mu} \text{ para } i_1 = 0, N_1, \tag{1}$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau/2} = \Lambda_1 y^{j+1/2} + \Lambda_2 y^{j+1} + \varphi^j, \quad x \in \omega_h,$$

$$y^{j+1} = \mu^{j+1} \text{ para } i_2 = 0, N_2, \quad y^0 = u_0(x), \quad x \in \bar{\omega}_h, \tag{2}$$

donde $\bar{\mu}$ es el valor intermedio de la función $\mu(x, t)$ igual a

$$\bar{\mu} = \frac{\mu^j + \mu^{j+1}}{2} - \frac{\tau}{4} \Lambda_2 (\mu^{j+1} - \mu^j).$$

Para hallar $y^{j+1/2}$ e y^{j+1} tenemos problemas de contorno en diferencias

$$\begin{gathered}
{}^1/_2 \tau \Lambda_1 y^{j+1/2} - y^{j+1/2} = -F^j, \\
F^j = y^j + 1/2\tau (\Lambda_2 y^j + \varphi^j), \quad x \in \omega_h, \\
y^{j+1/2} = \bar{\mu}, \quad i_1 = 0, N_1, \\
{}^1/_2 \tau \Lambda_2 y^{j+1} - y^{j+1} = -F^{j+1/2}, \\
F^{j+1/2} = y^{j+1/2} + {}^1/_2 \tau (\Lambda_1 y^{j+1/2} + \varphi^j), \quad x \in \omega_h, \\
y^{j+1} = \mu^{j+j}, \quad i_2 = 0, N_2.
\end{gathered} \tag{3}$$

El primer problema se resuelve mediante la factorización por las filas ($i_2 = 1, 2, \ldots, N_2 - 1$); el segundo, mediante la factorización por las columnas ($i_1 = 1, 2, \ldots, N_1 - 1$). El número de operaciones correspondientes a un nodo es finito y no depende de la red.

El esquema (3) es estable tanto respecto de los datos iniciales, como respecto del segundo miembro para cualesquiera τ y $|h|$ y tiene la exactitud $O(\tau^2 + |h|^2)$. De esto podemos convencernos eliminando $y^{j+1/2}$ y reduciendo el esquema (1), (2) a un esquema equivalente de dos capas con el operador factorizado B_j:

$$B\frac{y^{j+1}-y^j}{\tau} + Ay^j = \Phi^j, \quad j=0, 1, \ldots, \quad y^0 = u_0 \in H, \qquad (4)$$

$$B = \left(E + \frac{\tau}{2} A_1\right)\left(E + \frac{\tau}{2} A_2\right), \quad A_\alpha y = -\mathring{\Lambda}_\alpha y = -\mathring{y}_{\bar{x}_\alpha x_\alpha}, \quad \alpha = 1, 2,$$

donde $H = \Omega$ es el espacio de funciones reticulares definidas en los nodos interiores de la red ω_h.

Es evidente que $A_\alpha = A_\alpha^* > 0$, $\alpha = 1, 2$, $A_1A_2 = A_2A_1$. Por eso $B = E + \tau A/2 + \tau^2 A_1A_2/2 \geqslant E + \tau A/2 > \tau A/2$, y el esquema es estable.

2. Esquemas factorizados. El operador B, representado como un producto de varios operadores $B = B_1B_2 \ldots B_p$ se llamará *factorizado*, y el esquema correspondiente

$$B\frac{y^{j+1}-y^j}{\tau} + Ay^j = \varphi^j, \quad j=0, 1, \ldots, \quad y^0 = y(0), \qquad (5)$$

esquema factorizado.

Si para la resolución del problema

$$B_\alpha v = F_\alpha, \qquad \alpha = 1, 2,$$

con el segundo miembro prefijado F_α se requiere $O(N_1N_2)$ número de operaciones, entonces para determinar y^{j+1}, partiendo de y^j conocido, también son necesarias $O(N_1N_2)$ operaciones (el operador B es «económico)». Por cuanto

$$By^{j+1} = B_1B_2y^{j+1} = F^j,$$

el algoritmo se reducirá a la resolución sucesiva de las ecuaciones

$$B_1y^{j+1/2} = F^j, \qquad B_2y^{j+1} = y^{j+1/2}.$$

Apoyándose en la teoría de estabilidad de los esquemas de dos capas, no es difícil, partiendo del esquema con pesos, construir un esquema factorizado económico (por el método de regularización).

Así pues, supongamos que

$$A = A_1 + A_2, \quad B = E + \sigma\tau A = E + \sigma\tau (A_1 + A_2),$$
$$A_1 = A_1^*, \qquad A_2 = A_2^*.$$

En este caso el esguema (9) del § 2 será estable para $\sigma \geqslant \sigma_0 = \frac{1}{2} - \frac{1}{\tau \|A\|}$. Sustituyamos en (9) el operador B por un operador factorizado

$$\widetilde{B} = (E + \sigma\tau A_1)(E + \sigma\tau A_2),$$

que se diferencia de B en el término $\sigma^2\tau^2 A_1 A_2$,

$$\widetilde{B} = B + \sigma^2\tau^2 A_1 A_2.$$

Como resultado obtenemos un esquema factorizado

$$\widetilde{B}\,\frac{y^{j+1} - y^j}{\tau} + Ay^j = \Phi^j, \quad j = 0, 1, \ldots, \; y^0 = u_0 \in H, \tag{6}$$

del mismo orden de aproximación $O((\sigma - 1/2)\tau + \tau^2)$ que tiene el esquema de partida con pesos. Por cuanto el esquema de partida con pesos es estable ($\sigma \geqslant \sigma_0$), el esquema factorizado (6) será estable en virtud de la condición

$$\widetilde{B} > B > \tau A/2,$$

que se verifica, siempre que A_1 y A_2 son permutables y $A_\alpha^* = A_\alpha > 0$, $\alpha = 1, 2$.

Para hallar y^{j+1} obtenemos una ecuación $\widetilde{B}y^{j+1} = F^j$, o bien

$$(E + \sigma\tau A_1)(E + \sigma\tau A_2)\, y^{j+1} = F^j,$$
$$F^j = \widetilde{B}y^j + \tau(\Phi^j - Ay^j),$$

la cual se resuelve sucesivamente:

$$(E + \sigma\tau A_1)\,\bar{y} = F^j, \qquad (E + \sigma\tau A_2)\, y^{j+1} = \bar{y}$$

(con las condiciones de contorno correspondientes). El algoritmo que sigue es más económico (a cuenta del cálculo del segundo miembro F^j):

$$(E + \sigma\tau A_1)\, w^{j+1/2} = F^j = \Phi^j - Ay^j,$$
$$(E + \sigma\tau A_2)\, w^{j+1} = w^{j+1/2}, \quad y^{j+1} = y^j = \tau w^{j+1}. \tag{7}$$

Sin embargo, en este caso deben guardarse no uno, sino dos vectores ($w^{j+1/2}$ ó w^{j+1} e y^j). Cuando $\sigma = 1$, de (7) se deduce el segundo esquema de direcciones variables (*esquema de Duglass—Recford*)

$$\frac{y^{j+1/2}-y^j}{\tau}+A_1y^{j+1/2}+A_2y^j=\Phi^j,$$

$$(E+\tau A_2)\frac{y^{j+1}-y^{j+1/2}}{\tau}=\frac{y^{j+1/2}-y^j}{\tau}.$$

3. Método de aproximación sumaria. Con el fin de obtener esquemas económicos para la amplia clase de problemas (ecuaciones con coeficientes variables, dominios de forma compleja, etc.) hemos de cambiar el concepto de esquema de diferencias.

Dejamos a parte el concepto habitual de aproximación que se ha examinado anteriormente y lo cambiamos por un concepto más débil de *aproximación sumaria*. Aclaremos esto. Supongamos que el paso de una capa j a la otra $j+1$ se efectúa en varias etapas, en cada una de las cuales se utiliza el esquema corriente de dos capas que no aproxima la ecuación de partida y, no obstante, la suma de residuos para todo esquema intermedio

$$\psi=\sum_{\alpha=1}^{p}\psi\alpha \tag{8}$$

tiende a cero, cuando tiende a cero el paso τ según la variable t.

La idea del método de aproximación sumaria puede explicarse con un ejemplo del problema de Cauchy para la ecuación diferencial ordinaria

$$\frac{du}{dt}+au=f(t),\qquad t>0,\qquad u(0)=u_0, \tag{9}$$

donde $a>0$ es un número. Supongamos que

$$a=a_1+a_2,\quad a_1>0,\quad a_2>0,\quad f(t)=f_1(t)+f_2(t) \tag{10}$$

Es evidente que una representación de tal índole es siempre posible.

Introduzcamos una red $\omega_\tau = \{t_j = j\tau,\ j = 0, 1, \ldots\}$ y en cada paso (t_j, t_{j+1}) resolvemos sucesivamente (en lugar de (9)) dos ecuaciones

$$\frac{1}{2}\frac{dv_{(1)}}{dt}+a_1v_{(1)}=f_1(t), \quad t_j\leqslant t\leqslant t_{j+1/2}=t_j+\frac{\tau}{2},$$
$$\frac{1}{2}\frac{dv_{(2)}}{dt}+a_2v_{(2)}=f_2(t), \quad t_{j+1/2}\leqslant t\leqslant t_{j+1} \tag{11}$$

con los siguientes datos iniciales

$$v_{(1)}(t_j)=v(t_j), \quad v_{(2)}(t_{j+1/2})=v_{(1)}(t_{j+1/2}),$$
$$j=0, 1, \ldots, v_{(1)}(0)=u_0. \tag{12}$$

Como solución del problema (11)— 12) interviene la función

$$v(t)=v_{(2)}(t). \tag{13}$$

Cada una de las ecuaciones (11) se aproximará mediante un esquema de diferencias de dos capas con paso $\tau/2$. Por ejemplo, tomemos un esquema implícito

$$\frac{y^{j+1/2}-y^j}{\tau}+a_1y^{j+1/2}=f_1^j, \tag{14}$$
$$\frac{y^{j+1}-y^{j+1/2}}{\tau}+a_2y^{j+1}=f_2^j$$

Calculemos los residuos ψ_1 y ψ_2 para los esquemas (11). Sutituyamos en (11)

$$y^j=z^j+u^j, \quad y^{j+1/2}=z^{j+1/2}+u^{j+1/2}, \quad y^{j+1}=z^{j+1}+u^{j+1},$$
$$\frac{z^{j+1/2}-z^j}{\tau}+a_1z^{j+1/2}=-\psi_1^j,$$
$$\frac{z^{j+1}-z^{j+1/2}}{\tau}+a_2z^{j+1}=-\psi_2^j, \quad j=0, 1, \ldots,$$
$$z^0=0, \ \psi_1^j=\frac{u^{j+1/2}-u^j}{\tau}+a_1u^{j+1/2}-f_1^j,$$
$$\psi_2^j=\frac{u^{j+1}-u^{j+1/2}}{\tau}+a_2u^{j+1}-f_2^j.$$

Introduciendo aquí

$$u^{j+1}=(u+\tau\mathring{u}/2)^{j+1/2}+O(\tau^2), \quad u^j=(u-\tau\mathring{u}/2)^{j+1/2}+O(\tau^2)$$

obtenemos

$$\psi_1^j = (\dot{u}/2 + a_1 u - f_1)^{j+1/2} + O(\tau),$$
$$\psi_2^j = (\dot{u}/2 + a_2 u - f_2)^{j+1/2} + O(\tau). \qquad (15)$$

De aquí se ve que $\psi_1^j = O(1)$, $\psi_2^j = O(1)$, sin embargo

$$\psi_1^j + \psi_2^j = O(\tau) \to 0 \text{ cuando } \tau \to 0. \qquad (16)$$

Todos los razonamientos aducidos más arriba, a partir de (10), (11), (14), quedan en vigor si a_1 y a_2 representan las matrices o los operadores, y u, f, y son los vectores.

De este modo, el esquema (11), (12) aproxima el problema (9) en el sentido sumario (16) (tales esquemas se denominan *aditivos*).

Para demostrar la convergencia del esquema (11), (12) es menester obtener la estimación para el error $z^{j+1} = y^{j+1} - u^{j+1}$ en que se toma en consideración la propiedad (16) de la aproximación sumaria.

Pongamos

$$\psi_\alpha = \mathring{\psi}_\alpha + \psi_\varphi^*,$$
$$\mathring{\psi}_\alpha = (\mathring{u}/2 + a_\alpha u - f_\alpha)^{j+1/2}, \quad \psi_\alpha^* = O(\tau), \ \alpha = 1, 2,$$
$$z^{j+1/2} = \eta_{j+1/2} + \xi_{j+1/2}, \quad z^{j+1} = \eta_{j+1} + \xi_{j+1},$$

donde η_{j+1}, ξ_{j+1} son las soluciones de los problemas

$$\eta_{j+1/2} = \eta_j + \tau\mathring{\psi}_1, \quad \eta_{j+1} = \eta_{j+1/2} + \tau\mathring{\psi}_2,$$
$$j = 0, 1, \ldots, \ \eta_0 = 0, \qquad (17)$$

$$(1 + a_1\tau)\,\xi_{j+1/2} = \xi_j + \tau\widetilde{\psi}_1, \ (1 + a_2\tau)\,\xi_{j+1} = \xi_{j+1/2} + \tau\widetilde{\varphi}_2,$$
$$j = 0, 1, \ldots, \qquad (18)$$
$$\xi_0 = 0,$$

$$\widetilde{\psi}_1^j = \psi_1^{*j} - a_1\tau\eta_{j+1/2}, \quad \widetilde{\psi}_2^j = \psi_2^{*j} - a_2\tau\eta_{j+1}. \qquad (19)$$

De aquí encontramos $\eta_{j+1} = \eta_j + \tau\,(\mathring{\psi}_1^j + \mathring{\psi}_2^j) = \eta_j = \ldots = \eta_0 = 0$, es decir, $\eta_j = 0$ para cualquier $j = 0, 1, \ldots$ y $z^j = \xi_j$.

$$\eta_{j+1/2} = \tau\mathring{\psi}_1 = O(\tau), \qquad \widetilde{\psi}_\alpha = O(\tau). \qquad (20)$$

De (16) obtenemos

$$|\xi_{j+1/2}| \leqslant |\xi_j| + \tau |\widetilde{\psi}_1^j|,$$

$$|\xi_{j+1}| \leqslant |\xi_{j+1/2}| + \tau |\widetilde{\psi}_2^j| \leqslant |\xi_j| + \tau (|\widetilde{\psi}_1^j| + |\widetilde{\psi}_2^j|),$$

de modo que resulta lícita la estimación

$$|z^{j+1}| \leqslant \sum_{k=1}^{j} \tau (|\widetilde{\psi}_1^k| + /\widetilde{\psi}_2^k), \tag{21}$$

de la cual proviene precisamente (en virtud de (17)) la convergencia del sistema aditivo (14) con la velocidad $O(\tau)$.

En lugar de (11) podemos tomar otro sistema de ecuaciones:

$$\frac{dv_{(1)}}{dt} + a_1 v_{(1)} = f_1(t), \quad t_j \leqslant t \leqslant t_{j+1}, \; v_{(1)}(t_j) = v(t_j),$$

$$\frac{dv_{(2)}}{dt} + a_2 v_{(2)} = f_2(t), \; t_j \leqslant t \leqslant t_{j+1}, \; v_{(2)}(t_j) = v_{(1)}(t_{j+1}),$$

$$j = 0, \; 1, \; \ldots, \; v_{(1)}(0) = u_0.$$

Como solución de este problema interviene la función

$$v(t) = v_{(2)}(t). \tag{23}$$

A diferencia de (11), aquí ambas ecuaciones se interpretan en todo el segmento $t_j \leqslant t \leqslant t_{j+1}$, por lo cual la aproximación de dichas ecuaciones se realiza con el paso τ (y no con el paso $\tau/2$, como en el caso (11)) y da los mismos esquemas (14). Los dos métodos de reducción del problema (9) al sistema de problemas (11) ó (22) emplean una misma propiedad

$$a = a_1 + a_2 \tag{24}$$

y la condición $f = f_1 + f_2$, la cual siempre puede satisfacerse.

Veamos, como un ejemplo, la ecuación de conductibilidad térmica

$$\frac{\partial u}{\partial t} = Lu + f(x, t), \qquad x = (x_1, x_2), \tag{25}$$

$$Lu = \Delta u = L_1 u + L_2 u, \quad l_\alpha u = \frac{\partial^2 u}{\partial x_\alpha^2}, \quad \alpha = 1, 2,$$

L_1 y L_2 son los operadores «unidimensionales». La resolución de la ecuación

$$\frac{\partial v_{(\alpha)}}{\partial t} = L_\alpha v_{(\alpha)} + f_\alpha, \tag{26}$$

será, evidentemente, un problema más sencillo que la resolución de la ecuación (25). Las condiciones $L = L_1 + L_2$, $f = f_1 + f_2$ garantizan la aproximación sumaria para un esquema que se obtiene como resultado de la aproximación corriente, por ejemplo, con ayuda de un esquema de dos capas con pesos de cada una de las ecuaciones del sistema

$$\frac{dv_{(1)}}{dt} = L_1 v_{(1)} + f_1, \quad t_j \leqslant t \leqslant t_{j+1}, \quad v_{(1)}^j = v^j,$$

$$\frac{dv_{(2)}}{dt} = L_2 v_{(2)} + f_2, \quad t_j \leqslant t \leqslant t_{j+1}, \quad v_{(2)}^j = v_{(1)}^{j+1}, \quad v^{j+1} = v_{(2)}^{j+1}.$$

De resultas obtenemos un esquema aditivo, un esquema unidimensional local o bien un esquema de fisión

$$\frac{y^{j+1/2} - y^j}{\tau} = \Lambda_1 (\sigma_1 y^{j+1/2} + (1 - \sigma_1) y^j) + \varphi_1^j, \quad x \in \omega_h,$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau} = \Lambda_2 (\sigma_2 y^{j+1} + (1 - \sigma_2) y^{j+1/2}) + \varphi_2^j,$$

$$x \in \omega_h, \ j = 0, 1, \ldots, \tag{27}$$

$$y^0 = u_0(x), \qquad x \in \omega_h,$$

$$y^{j+1/2}|_{\gamma_h} = \mu^{j+1/2}, \qquad y^{j+1}|_{\gamma_h} = \mu^{j+1}.$$

Aquí $\Lambda_1 y = y_{\bar{x}_1 x_1}$, $\Lambda_2 y = y_{\bar{x}_2 x_2}$. Los parámetros σ_1 y σ_2 se determinan partiendo de las condiciones de estabilidad y aproximación. Por ejemplo, cuando $\sigma_1 = \sigma_2 = 1$, obtenemos un esquema con adelanto

$$\frac{y^{j+1/2} - y^j}{\tau} = \Lambda_1 y^{j+1/2} + \varphi_1^j,$$

$$\frac{y^{j+1} - y^{j+1/2}}{\tau} = \Lambda_2 y^{j+1} + \varphi_2^j, \ j = 0, 1, \ldots$$

Al sustituir aquí $y^j = z^j + u^j$, $y^{j+1/2} = z^{j+1/2} + (u^j + u^{j+1})/2$, $y^{j+1} = z^{j+1} + u^{j+1}$, obtenemos para el error z

las ecuaciones

$$\frac{z^{j+1/2}-z^j}{\tau}=\Lambda_1 z^{j+1/2}+\psi_1^j,$$
$$\frac{z^{j+1}-z^{j+1/2}}{\tau}=\Lambda_2 z^{j+1}+\psi_2^j,$$

donde u es la solución del problema de partida (25), ψ_1 y ψ_2 son los residuos

$$\psi_1^j=\Lambda_1\frac{u+\hat{u}}{2}-\frac{1}{2}\frac{\hat{u}-u}{\tau}+\varphi_1,\quad \psi_2^j=\Lambda_2\hat{u}-\frac{1}{2}\frac{\hat{u}-u}{\tau}+\varphi_2,$$
$$\hat{u}=u^{j+1},\qquad u=u^j.$$

De aquí se ve que $\psi_1=O(1)$, $\psi_2=O(1)$, es decir, cada una de las ecuaciones (27) no aproxima, tomada separadamente, la ecuación (25). Tomemos la suma de residuos

$$\psi=\psi_1+\psi_2=\Lambda_1\frac{u+\hat{u}}{2}+\Lambda_2\hat{u}-\frac{\hat{u}-u}{\tau}+\varphi_1+\varphi_2=$$
$$=(L_1+L_2)\bar{u}-\frac{\partial\bar{u}}{\partial t}+\varphi_1+\varphi_2+O(\tau+|h|^2),$$

donde $\bar{u}=u^{j+1/2}$. Tomando en consideración la ecuación (25) para $t=t_{j+1/2}$, obtendremos

$$\psi=\varphi_1+\varphi_2-f^{j+1/2}+O(\tau+|h|^2)=O(\tau+|h|^2)$$
$$|h|^2=h_1^2+h_2^2,$$

siempre que

$$\varphi_1+\varphi_2=f^{j+1/2}+O(\tau^2).$$

Esto puede ser conseguido suponiendo, por ejemplo,

$$\varphi_1=0,\ \varphi_2=f^{j+1/2}\quad \text{o bien}\quad \varphi_1=\varphi_2=f^j/2.$$

Se puede mostrar que el esquema (27) converge uniformemente con la velocidad

$$O(\tau+|h|^2),\ \text{es decir},\ \|y^{j+1}-u^{j+1}\|_C=O(\tau+|h|^2).$$

De los ejemplos aducidos se ve que el método de aproximación sumaria permite realizar la partición de los problemas complejos en una sucesión de problemas más sencillos y simplificar considerablemente la resolución de los problemas multidimensionales de la física matemática.

Anexo

Algoritmo de marcha y método de reducción para resolver sistemas de ecuaciones lineales con matriz tridiagonal

En varias aplicaciones se encuentran problemas que conducen a la resolución de los sistemas de ecuaciones algebraicas lineales especiales (con una matriz enrarecida que cuenta con muchos elementos nulos) de orden superior. Los sistemas de tal género surgen al realizar una aproximación de diferencias de las ecuaciones elípticas o bien al utilizar esquemas implícitos para la ecuación de conductibilidad térmica.

Al aproximar en el cap. IV una ecuación diferencial corriente de segundo orden en el molde tripuntual, hemos obtenido una ecuación en diferencias de segundo orden la cual representa un sistema de ecuaciones algebraicas lineales de $(N-1)$-ésimo orden ($N-1$ es el número de nodos interiores) con la matriz tridiagonal. Con el objeto de resolver dicho sistema se ha construido en el § 3, cap. I un método cuya realización requiere $O(N)$ operaciones aritméticas.

En el cap. VI hemos obtenido, aproximando la ecuación de Poisson bidimensional en un molde pentapuntual, un esquema de diferencias, a la cual corresponde el sistema de ecuaciones algebraicas lineales con matriz pentadiagonal de orden $N=(N_1-1)(N_2-1)$, donde N_1-1, N_2-1 es el número de nodos interiores según cada dirección. Al partir el vector de incógnitas en bloques, cada uno de los cuales contenga N_1-1 elementos, obtendremos una inscripción del sistema con matriz tridiagonal de bloques, con la particularidad de que el número de bloques en la matriz citada es igual a N_2-1. Para tal sistema hemos estudiado en el § 2, cap. VI el método de separación de variables con la estimación $O(N \log N)$ para el número de operaciones. Cuando los sistemas semejantes se resuelven varias veces se hace muy importante que los algoritmos computacionales sean económicos.

Más abajo construiremos un método directo para resolver sistemas especiales con la matriz triagonal, el cual exige sólo $O(N)$ operaciones tanto en el caso en que los elementos de la matriz son escalares, como en el caso de la matriz de bloques.

1. Estudiemos al principio un caso en que los elementos de la matriz son escalares. Escribamos el sistema con una matriz tridiagonal en forma de un problema de diferencias tripuntual:

$$-y_{i-1}+Cy_i-y_{i+1}=F_i, \quad 1 \leqslant i \leqslant N-1, \quad y_0=0, \quad y_N=0, \qquad (1)$$

donde C es un número, y supongamos que $N=2k+1$. Si escribimos la ecuación en diferencias de segundo orden (1) en forma de las rela-

ciones recurrentes

$$y_{i+1} = Cy_i - y_{i-1} - F_i, \quad i \geqslant 1, \quad y_0 = 0, \tag{2}$$

no será difícil notar que todas las incógnitas y_i pueden ser encontradas sucesivamente por la fórmula (2), si calculamos y_1 de tal o cual modo. En este caso, cualquier y_i se expresará linealmente en términos de y_0 e y_1. Todo lo dicho nos permite escribir para cualquier $i \geqslant 1$ una correlación

$$y_{i+1} = \alpha_i y_1 - \beta_{i-1} y_0 - p_i \tag{3}$$

con los coeficientes α_i, β_i, p_i que por ahora quedan indeterminados. Si ponemos

$$\alpha_0 = 1, \quad \beta_{-1} = 0, \quad p_0 = 0, \tag{4}$$

entonces (3) se verifica también para $i = 0$. Así pues, la solución del problema se buscará en la forma (3) para cualquier $i \geqslant 0$.

Anotando (1) en forma de las relaciones recurrentes

$$y_{i-1} = Cy_i - y_{i+1} - F_i, \quad i \leqslant N - 1, \quad y_N = 0 \tag{5}$$

y razonando análogamente, llegamos a que la solución del problema (1) para cualquier $i \leqslant N$ se puede buscar en la forma

$$y_{i-1} = \xi_{N-i} y_{N-1} - \eta_{N-i-1} y_N - q_{N-i}, \tag{6}$$

si ponemos

$$\xi_0 = 1, \quad \eta_{-1} = 0, \quad q_0 = 0. \tag{7}$$

Observemos que si y_{N-1} queda determinado, todos los y_i se pueden calcular sucesivamente según la fórmula (5).

Hallemos y_1 e y_{N-1}. Con este fin determinemos los coeficientes α_i, β_i, ξ_i, η_i, p_i, q_i. Al comparar (2) y (3) para $i = 1$, y (5) y (6), para $i = N - 1$, obtendremos

$$\alpha_1 = \xi_1 = C, \quad \beta_0 = \eta_0 = 1, \quad p_1 = F_1, \quad q_1 = F_{n-1}. \tag{8}$$

Encontremos ahora las fórmulas recurrentes para determinar los coeficientes buscados. Sustituyamos (3), al igual que las expresiones para y_i e y_{i-1} que se desprenden de (3):

$$y_i = \alpha_{i-1} y_1 - \beta_{i-2} y_0 - p_{i-1}, \quad y_{i-1} = \alpha_{i-2} y_1 - \beta_{i-3} y_0 - p_{i-2},$$

en la ecuación (1). Obtendremos

$$-(\alpha_{i-2} - C\alpha_{i-1} + \alpha_i)\, y_1 + (\beta_{i-3} - C\beta_{i-2} + \beta_{i-1})\, y_0 + \\ + p_{i-2} - Cp_{i-1} + p_i = F_i, \quad i \geqslant 2.$$

Para que estas igualdades sean idénticas con i cualquiera es suficiente hacer para $i \geqslant 2$

$$p_i = Cp_{i-1} - p_{i-2} + F_i, \tag{9}$$

$$\alpha_i = C\alpha_{i-1} - \alpha_{i-2}, \quad \beta_{i-1} = C\beta_{i-2} - \beta_{i-3}. \tag{10}$$

Análogamente, haciendo uso de (6) y (1), obtendremos para $i \leqslant N-2$ las relaciones recurrentes

$$q_{N-i}=Cq_{N-i-1}-q_{N-i-2}+F_i,$$

$$\xi_{N-i}=C\xi_{N-i-1}-\xi_{N-i-1},\quad \eta_{N-i-1}=C\eta_{N-i-2}-\eta_{N-i-3}.$$

Al sustituir aquí $N-i$ por i, tenemos para $i \geqslant 2$ las fórmulas siguientes

$$q_i = Cq_{i-1} - q_{i-2} + F_{N-i}, \tag{11}$$

$$\xi_i = C\xi_{i-1} - \xi_{i-2},\quad \eta_{i-1} = C\eta_{i-2} - \eta_{i-3}. \tag{12}$$

Así pues, las fórmulas (4), (7)—(12) determinan por completo los coeficientes buscados. Al cotejar (10) y (12) bajo las condiciones (4), (7), (8), llegamos a que $\beta_i = \eta_i = \xi_i = \alpha_i$ para $i \geqslant 0$. De este modo, las fórmulas (3), (6) toman la forma

$$y_{i+1} = \alpha_i y_1 - \alpha_{i-1} y_0 - p_i,\quad i \geqslant 0, \tag{13}$$

$$y_{i-1} = \alpha_{N-i} y_{N-1} - \alpha_{N-i-1} y_N - q_{N-i},\quad i \leqslant N, \tag{14}$$

donde

$$p_i = Cp_{i-1} - p_{i-2} + F_i,\quad i \geqslant 2,\quad p_0 = 0,\quad p_1 = F_1, \tag{15}$$

$$q_i = Cq_{i-1} - q_{i-2} + F_{N-i},\quad i \geqslant 2,\quad q_0 = 0,\quad q_1 = F_{N-1} \tag{16}$$

$$\alpha_i = C\alpha_{i-1} - \alpha_{i-2},\quad i \geqslant 2,\quad \alpha_0 = 1,\quad \alpha_1 = C. \tag{17}$$

Hallemos ahora y_1 e y_{N-1}. Con este fin pongamos en (13) $i = k$, y en (14) $i = k + 2$. Teniendo presente que $N = 2k + 1$, tendremos

$$y_{k+1} = \alpha_k y_1 - \alpha_{k-1} y_0 - p_k,\qquad y_{k+1} = \alpha_{k-1} y_{N-1} - \alpha_{k-2} y_N - q_{k-1}.$$

Restando la primera igualdad de la segunda, obtendremos una ecuación respecto de y_1 e y_{N-1}:

$$\alpha_{k-1} y_{N-1} - \alpha_k y_1 + \alpha_{k-1} y_0 - \alpha_{k-2} y_N = q_{k-1} - p_k. \tag{18}$$

Obtengamos una ecuación más para y_1 e y_{N-1}, suponiendo $i = k - 1$ en (13) y $i = k + 1$ en (14), y sustrayendo la segunda ecuación de la primera

$$-\alpha_k y_{N-1} + \alpha_{k-1} y_1 - \alpha_{k-1} y_0 + \alpha_{k-1} y_N = p_{k-1} - q_k. \tag{19}$$

Teniendo presente que $y_0 = y_N = 0$, sumemos y sustrayamos (18) y (19). Obtendremos un esquema equivalente

$$\begin{aligned}(\alpha_{k-1} - \alpha_k)\,(y_{N-1} + y_1) &= q_{k-1} - p_k + p_{k-1} - q_k,\\ (\alpha_{k-1} + \alpha_k)\,(y_{N-1} - y_1) &= q_{k-1} - p_k - p_{k-1} + q_k;\end{aligned} \tag{20}$$

al resolver dicho sistema hallaremos los valores buscados de y_1 e y_{N-1}:

$$\begin{aligned}y_1 &= (\alpha_{k-1}^2 - \alpha_k^2)^{-1}\,[\alpha_k\,(q_{k-1} - p_k) + \alpha_{k-1}\,(p_{k-1} - q_k)],\\ y_{N-1} &= (\alpha_{k-1}^2 - \alpha_k^2)^{-1}\,[\alpha_{k-1}\,(q_{k-1} - p_k) + \alpha_k\,(p_{k-1} - q_k)].\end{aligned} \tag{21}$$

De este modo, el algoritmo de resolución del problema (1) consiste en calcular por las fórmulas (15)—(17) los coeficientes p_{k-1}, p_k, q_{k-1}, q_k, α_{k-1}, α_k, por las formulas (21) los valores de y_1, y_{N-1}, por la fórmula (2) las incógnitas y_i, $i = 2, 3, \ldots, k$ y por la fórmula (5) para $i = N-2, N-3, \ldots, k+1$ con y_0, y_N dados e y_1, y_{N-1} calculados. El algoritmo descrito recibió el nombre de *algoritmo de marcha*. Es fácil calcular que para su realización se necesitan aproximadamente $8N$ operaciones. Podemos mostrar que si $C \neq 2 \cos m\pi/N$, m es un número entero, entonces el problema (1) es resoluble para cualquier miembro segundo y $\alpha_{k-1}^2 \neq \alpha_k^2$. Por consiguiente en este caso las fórmulas (21) no contienen la operación de división por cero.

El algoritmo de marcha descrito arriba puede ser empleado también en un caso en que C es una matriz cuadrada, F_1 son los vectores prefijados, e y_i, los vectores buscados. Ha de notarse que el problema de Dirichlet de diferencias para la ecuación de Poisson (véase el cap. VI) sobre una red rectangular uniforme según cualquier dirección, introducida en el rectángulo, puede ser escrito en la forma (1). En este caso los valores de la función reticular buscada correspondientes a la i-ésima fila son los componentes del vector, mientras que la matriz C es tridiagonal y su orden es igual al número de filas interiores de la red.

Sea M el orden de la matriz C. Entonces los vectores p_i, q_i son de dimensión M y para calcular p_{k-1}, q_{k-1}, p_k, q_k según las fórmulas (15), (16) se exigirán $O(MN)$ operaciones. Es evidente que el mismo número de operaciones se necesitarán también para encontrar los vectores y_i, $2 \leqslant i \leqslant N-2$ según las fórmulas (2), (5). Veamos ahora la cuestión sobre el cálculo de y_1 e y_{N-1}.

De la fórmula (17) se deduce que α_k es un polinomio de grado k de C, además, si C es un número, entonces α_k será un polinomio algebraico, y si C es una matriz, α_k será un polinomio matricial. Para un polinomio que satisface la relación recurrente (17) existe una representación explícita: $\alpha_k = U_k(C/2)$, donde $U_k(x)$ es el polinomio de Chébishev de segunda especie de grado k:

$$U_k(x) = \begin{cases} \dfrac{\operatorname{sen}(k+1)\arccos x}{\operatorname{sen}\arccos x}, & |x| \leqslant 1, \\[2ex] \dfrac{\left(x+\sqrt{x^2-1}\right)^{k+1} - \left(x-\sqrt{x^2-1}\right)^{k+1}}{2\sqrt{x^2-1}}, & |x| \geqslant 1. \end{cases}$$

Haciendo uso de la expresión explícita para α_k, $k \geqslant 0$, y teniendo presente que α_k es un polinomio cuyo grado mayor tiene el coeficiente unidad, podemos obtener los siguientes desarrollos:

$$\begin{aligned} \alpha_k - \alpha_{k-1} &= \prod_{l=1}^{k} \left(C - 2\cos\frac{(2l-1)\pi}{2k+1} E \right), \\ \alpha_k + \alpha_{k-1} &= \prod_{l=1}^{k} \left(C - 2\cos\frac{2l\pi}{2k+1} E \right). \end{aligned} \qquad (22)$$

Recurriendo a (22) y (23), construyamos el siguiente algoritmo para hallar y_1 e y_{N-1}:

$$v_0 = p_k - q_{k-1} - p_{k-1} + q_k, \quad w_0 = q_{k-1} - p_k - p_{k-1} + q_k,$$

$$\left(C - 2\cos\frac{(2l-1)\pi}{2k+1}E\right) v_l = v_{l-1},$$

$$\left(C - 2\cos\frac{2l\pi}{2k+1}E\right) w_l = w_{l-1}, \quad l-1, 2, \ldots, k, \tag{23}$$

$$y_1 = 0{,}5\,(v_k - w_k), \quad y_{N-1} = 0{,}5\,(v_k + w_k).$$

Por cuanto cada uno de los sistemas (23) cuenta con una matriz tridiagonal (el número de tales sistemas es $2k$) y puede ser resuelto por el método de factorización realizando $O(M)$ operaciones, entonces para encontrar y_1 e y_{N-1} se necesitarán $O(NM)$ operaciones aritméticas.

Así pues, para resolver el sistema (1) con la matriz triagonal se ha construido un método en el que el número de operaciones aritméticas es proporcional al número de incógnitas.

Fijémonos en que el algoritmo de marcha construido puede ser numéricamente inestable. En efecto, si el número C satisface la condición $|C| > 2$, entonces para el algoritmo resulta característico el crecimiento del error, exponencial según N, puesto que entre las raíces de la ecuación característica $q^2 - Cq + 1 = 0$ se tiene una que en módulo es superior a la unidad. La inestabilidad del mismo tipo tiene lugar también cuando la matriz C tiene valores propios que son superiores en módulo a 2. Para los problemas de esta índole está construida actualmente una variante del algoritmo de marcha estable en aquel sentido que el error crece, al crecer N, según la ley potencial.

2. **Método de reducción.** En algunos casos, al resolver los sistemas de ecuaciones algebraicas lineales con una matriz tridiagonal, es de mucha importancia la exactitud de la solución obtenida. El análisis de las fórmulas del método de factorización que se emplea para resolver los sistemas citados muestra que la fuente del error puede radicar en las fórmulas para calcular los coeficientes de factorización. Estas fórmulas contienen la operación de división por una diferencia de las magnitudes que son próximas en su valor. Más abajo se dará a conocer el método de reducción para resolver dichos sistemas, privado de la deficiencia mencionada.

Así pues, supongamos que se requiere hallar la solución de un problema de diferencias tripuntual

$$-a_i y_{i-1} + c_i y_i - b_i y_{i+1} = f_i, \quad 1 \leqslant i \leqslant N-1, \tag{24}$$

$$y_0 = 0, \quad y_N = 0,$$

donde $c_i = a_i + b_i + d_i$, $a_i > 0$, $b_i > 0$, $d_i \geqslant 0$, $N = 2^n$. La idea del método de reducción consiste en que del sistema (24) se eliminan consecutivamente las incógnitas con números impares primeramente, y a continuación con los números múltiplos de 2, etc.

Escribamos tres ecuaciones del sistema (24) que vienen una tras otra con los números $i-1$, i, $i+1$, donde i es un número par.

$$-a_{i-1}y_{i-2}+(a_{i-1}+b_{i-1}+d_{i-1})\,y_{i-1}-b_{i-1}y_i=f_{i-1}, \tag{25}$$

$$-a_iy_{i-1}+(a_i+b_i+d_i)\,y_i-b_iy_{i+1}=f_i, \tag{26}$$

$$-a_{i+1}y_i+(a_{i+1}+b_{i+1}+d_{i+1})\,y_{i+1}-b_{i+1}y_{i+2}=f_{i+1}. \tag{27}$$

Multiplicando la ecuación (25) por $\alpha_i^{(1)}=a_i\,(a_{i-1}+b_{i-1}+d_{i-1})^{-1}$, la ecuación (27) por $\beta_i^{(1)}=b_i\,(a_{i+1}+b_{i+1}+d_{i+1})^{-1}$ y sumando las ecuaciones obtenidas con (26) llegamos a que

$$-a_i^{(1)}y_{i-2}+(a_i^{(1)}+b_i^{(1)}+d_i^{(1)})\,y_i-b_i^{(1)}y_{i+2}=f_i^{(1)},$$

$$i=2,\ 4,\ 6,\ \ldots,\ N-2,\quad y_0=0,\quad y_N=0, \tag{28}$$

donde $a_i^{(1)}=\alpha_i^{(1)}a_{i-1}$, $b_i^{(1)}=\beta_i^{(1)}b_{i+1}$, $d_i^{(1)}=\alpha_i^{(1)}d_{i-1}+d_i+$ $+\beta_i^{(1)}d_{i+1}$, $f_i^{(1)}=\alpha_i^{(1)}f_{i-1}+f_i+\beta_i^{(1)}f_{i+1}$. Si las incógnitas con números pares quedan halladas (ellas satisfacen el sistema (28)), entonces las demás incógnitas se determinarán según la fórmula

$$y_i=\frac{f_i+a_iy_{i-1}+b_iy_{i+1}}{a_i+b_i+d_i},\quad i=1,\ 3,\ 5,\ \ldots,\ N-1,$$

Es evidente que el proceso descrito de eliminación de las incógnitas puede ser aplicado al sistema (28), del cual serán eliminadas en el segundo paso las incógnitas con los números múltiplos de 2, pero no múltiplos de 4. Como resultado del l-ésimo paso del proceso de eliminación obtendremos un sistema

$$-a_i^{(l)}y_{i-2^l}+(a_i^{(l)}+b_i^{(l)}+d_i^{(l)})\,y_i-b_i^{(l)}y_{i+2^l}=f_i^{(l)}, \tag{29}$$

$$i=2^l,\quad 2\cdot2^l,\quad 3\cdot2^l,\ \ldots,\ N-2^l,\ y_0=0,\ y_N=0,$$

donde

$$\begin{gathered}
a_i^{(l)}=\alpha_i^{(l)}a_{i-2^{l-1}}^{(l-1)},\quad b_i^{(l)}=\beta_i^{(l)}b_{i+2^{l-1}}^{(l-1)},\\
d_i^{(l)}=\alpha_i^{(l)}d_{i-2^{l-1}}^{(l-1)}+d_i^{(l-1)}+\beta_i^{(l)}d_{i+2^{l-1}}^{(l-1)},\\
f_i^{(l)}=\alpha_i^{(l)}f_{i-2^{l-1}}^{(l-1)}+f_i^{(l-1)}+\beta_i^{(l)}f_{i+2^{l-1}}^{(l-1)},\\
\alpha_i^{(l)}=a_i^{(l-1)}\left(a_{i-2^{l-1}}^{(l-1)}+b_{i-2^{l-1}}^{(l-1)}+d_{i-2^{l-1}}^{(l-1)}\right)^{-1},\\
\beta_i^{(l)}=b_i^{(l-1)}\left(a_{i+2^{l-1}}^{(l-1)}+b_{i+2^{l-1}}^{(l-1)}+d_{i+2^{l-1}}^{(l-1)}\right)^{-1},\\
i=2^l,\quad 2\cdot2^l,\quad 3\cdot2^l,\ \ldots,\ N-2^l,\quad l\geqslant1.
\end{gathered} \tag{30}$$

Aquí se usan las designaciones $a_i^{(0)}=a_i$, $b_i^{(0)}=b_i$, $d_i^{(0)}=d_i$, $f_i^{(0)}=f_i$.

El proceso de eliminación se dará por terminado en el $(n-1)$-ésimo paso, cuando el sistema (29) se componga de una sola ecuación

respecto de la incógnita $y_{N/2} = y_{2^{n-1}}$. De esta ecuación encontraremos

$$y_{2^{n-1}} = \frac{f_{2^{n-1}}^{(n-1)} + a_{2^{n-1}}^{(n-1)} y_0 + b_{2^{n-1}}^{(n-1)} y_N}{a_{2^{n-1}}^{(n-1)} + b_{2^{n-1}}^{(n-1)} + d_{2^{n-1}}^{(n-1)}}, \quad y_0 = y_N = 0. \tag{31}$$

Las incógnitas restantes se determinarán según las fórmulas

$$y_i = \frac{f_i^{(l)} + a_i^{(l)} y_{i-2^l} + b_i^{(l)} y_{i+2^l}}{a_i^{(l)} + b_i^{(l)} + d_i^{(l)}}, \quad i = 2^l, \quad 3 \cdot 2^l,$$

$$5 \cdot 2^l, \ \ldots, \ N - 2^l, \tag{32}$$

donde $l = n - 2, \ n - 3, \ \ldots, \ 0, \ y_0 = y_N = 0$. Observemos que la fórmula (32) incluye la fórmula (31) para $l = n - 1$.

Así pues, aplicándose el método de reducción, en el paso directo se calculan $a_i^{(l)}$, $b_i^{(l)}$, $d_i^{(l)}$, $f_i^{(l)}$ según las fórmulas (30) para $l = 1, 2, \ldots, n - 1$, y en el paso inverso se halla la solución buscada por la fórmula (32) para $l = n - 1, n - 2, \ldots, 0$. Ha de ser notado que el método no requiere una memoria complementaria, puesto que las magnitudes $a_i^{(l)}$, $b_i^{(l)}$, $d_i^{(l)}$, $f_i^{(l)}$ pueden ser dispuestas en los lugares respectivos de $a_{i-2^{l-1}}^{(l-1)}$, $b_{i-2^{l-1}}^{(l-1)}$, $d_i^{(l-1)}$, $f_i^{(l-1)}$. Para realizar el método se necesitan $12N$ adiciones, $8N$ multiplicaciones y $3N$ divisiones.

Bibliografía

1. Bakhválov N. S. Numérical Methods, Ed. Mir, M., 1978
2. Березин И. С., Жидков Н. П. Методы вычислений. М., Наука, 1966, ч. I; Физматгиз, 1962, ч. 2
(Berézin I. S., Zhidkov N. P. Métodos de los cálculos).
3. Воеводин В. В. Численные методы алгебры; теория и алгоритмы. М., Наука, 1966
(Voevodin V. V. Métodos numéricos del álgebra; teoría y algoritmos).
4. Годунов С. К., Рябенький В. С. Разностные схемы. М., Наука, 1977
(Godunov S. K., Riábenki V. S. Esquemas de diferencias).
5. Калиткин Н. Н. Численные методы. М., Наука, 1978
(Kalitkin N. N. Métodos numéricos).
6. Ляшко И. И., Макаров В. Л., Скоробогатько А. А. Методы вычислений. Киев, Высшая школа, 1977
(Liashkó I. I., Makárov V. L., Skorobogatko A. A. Métodos de los cálculos).
7. Марчук Г. И. Методы вычислительной математики. М., Наука, 1977
(Marchuk G. I. Métodos de las matemáticas de cálculo).
8. Никольский С. М. Квадратурные формулы. М., Наука, 1979
(Nikolski S. M. Fórmulas de cuadratura).
9. Самарский А. А. Теория разностных схем. М., Наука, 1977
(Samarski A. A. Teoría de los esquemas de diferencias).
10. Самарский А. А., Андреев В. Б. Разностные методы для эллиптических уравнений. М., Наука, 1976
(Samarski A. A., Andréiev V. B. Métodos de diferencias para las ecuaciones elípticas).
11. Самарский А. А., Гулин А. В. Устойчивость разностных схем. М., Наука, 1973
(Samarski A. A., Gulin A. V. Estabilidad de los esquemas de diferencias).
12. Samarski A. A., Nikoláev E. S. Métodos de solución de las ecuaciones reticulares, Ed. Mir, M., 1983.
13. Самарский А. А., Попов Ю. П. Разностные методы газовой динамики. М., Наука, 1980
(Samarski A. A., Popov Yu. P. Métodos de diferencias de la dinámica de los gases).

14. Tíjonov A. N., Samarski A. A. Ecuaciones de la física matemática, Ed. Mir, M., 1983.
15. Фаддеев Д. К., Фаддеева В. Н. Вычислительные методы линейной алгебры. М., Наука, 1972
(Faddéiev D. K., Faddéieva V. N. Métodos computacionales del álgebra lineal).
16. Яненко Н. Н. Метод дробных шагов решения многомерных задач математической физики. Новосибирск, Наука, 1967
(Yanenko N. N. Método de pasos fraccionarios para resolver problemas multidimensionales de la física matemática).

Lista de designaciones

$\omega_N = \{i;\ i = 0, 1, \ldots, N\}$, red con nodos de números enteros
$\overline{\omega}_h = \{x_i = ih,\ h = 1/N,\ 0 \leqslant i \leqslant N\}$, red uniforme de paso h en el segmento $[0, 1]$.
h, paso de la red ω_h
$y_i = y(x_i) = y(i)$, valor de la función reticular en el i-ésimo nodo de la red
$\hat{\omega}_h$, red no uniforme
$h_i = x_i - x_{i-1}$, paso de la red no uniforme $\hat{\omega}_h$:
$\hbar_i = \frac{1}{2}(h_i + h_{i+1})$
$v_{i_1 i_2} = v(x_{i_1}^1, x_{i_2}^2)$, valor de la función reticular bidimensional en el) nodo (i, j)
$v^n_{i_1 i_2} = v(x_{i_1}^1, x_{i_2}^2, t_n)$, valor de la función reticular en el nodo (i, j) en la n-ésima capa temporal
$v_{ij}^{n+1} = \hat{v}$, valor de la función reticular bidimensional en el nodo (i, j) sobre la $(n + 1)$-ésima capa
$\Delta y_i = y_{i+1} - y_i$, diferencia derecha en el i-ésimo nodo
$\nabla y_i = y_i - y_{i-1}$, diferencia izquierda en el i-ésimo nodo
$\delta y_i = \frac{1}{2}(\nabla y_i + \Delta y_i)$, diferencia central en el i-ésimo nodo
$\Delta^2 y_{i+1} = \Delta(y_{i+1}) = \Delta(\Delta y_i)$, diferencia de segundo orden
$y_{x,i} = (y_{i+1} - y_i)/h$, derivada de diferencias derecha en el nodo x_i
$y_{\bar{x},i} = (y_i - y_{i-1})/h$, derivada de diferencias izquierda en el nodo x_i
$y_{\overset{\circ}{x}, i} = (y_{i+1} - y_{i-1})/(2h)$, derivada de diferencias central en el nodo x_i
$y_{\bar{x}x,i} = (y_{i+1} - 2y_i + y_{i-1})/h^2$, segunda derivada de diferencias
H, espacio de Hilbert
(y, v), producto escalar de los elementos $y, v \in H$, $\| y \| = \sqrt{(y, y)}$
E, operador unidad
A^*, operador conjugado del operador A
A^{-1}, operador inverso al operador A
$A > 0$, operador positivo
$A \geqslant 0$, operador no negativo
$A \geqslant \delta E$, $\delta > 0$, operador definido positivo
$\| y \|_A = \sqrt{(Ay, y)}$, $y \in H$, norma energética
Espacio de funciones reticulares:

$\Omega_{N+1} = \{y_i,\ i = 0, \ldots, N\}$
$\overset{\circ}{\Omega}_{N-1} = \{y_i,\ i = 0, \ldots, N;\ y_0 = 0,\ y_N = 0\}$
$\overset{\circ}{y}_i$, función del $\overset{\circ}{\Omega}_{N+1}$
$\Omega_N^{*} = \{y_i,\ i = 0, 1, \ldots, N - 1\}$
$\Omega_N^{-} = \{y_i,\ i = 1, 2, \ldots, N\}$
Productos escalares y normas en la red:

$$(y, v) = \sum_{i=1}^{N-1} y_i v_i h, \quad \| y \| = \sqrt{(y, y)}$$

$$(y, v] = \sum_{i=1}^{N} y_i v_i h, \quad \| y] | = \sqrt{(y, y]}$$

$$\| y \|_C = \max_{x_i \in \bar{\omega}_h} | y(x_i) | = \max_{0 \leqslant i \leqslant N} | y(x_i) |$$

Indice alfabetico

A nuetros lectores:

«Mir» edita libros soviéticos traducidos al español, inglés, francés, árabe y otros idiomas extranjeros. Entre ellos figuran las mejores obras de las distintas ramas de la ciencia y la técnica: manuales para los centros de enseñanza superior y escuelas tecnológicas; literatura sobre ciencias naturales y médicas. También se incluyen monografías, libros de divulgación científica y ciencia-ficción.

Dirijan sus opiniones a la Editorial Mir, 1 Rizhski per., 2, 129820, Moscú, 1-110, GSP, URSS.

УДК 519.6=60

Александр Андреевич Самарский

ВВЕДЕНИЕ В ЧИСЛЕННЫЕ МЕТОДЫ

Контрольный редактор С. Я. Калашник
Редактор Н. А. Стальнова
Художники С. А. Бычков, Г. В. Чучелов
Художественный редактор Е. И. Подмарькова
Технический редактор В. П. Сизова
Корректор Г. А. Макарова

ИБ № 5146

Сдано в набор 22.05.85. Подписано к печати 19.12.85.
Формат 84 × 108¹/₃₂. *Бумага типографская* № 1.
Гарнитура обыкновенная. Печать высокая.
Объем 4,88 бум. л. Усл. печ. л. 15,05. Усл. кр.-отт. 16,92.
Уч.-изд. л. 15,05. Изд. № 19/3488. Тираж 13 380 экз.
Зак. 0242. Цена 1 р. 55 к.

ИЗДАТЕЛЬСТВО «МИР»
129820, Москва, И-110, ГСП, 1-й Рижский пер., 2.

Ордена Трудового Красного Знамени
Московская типография № 7 «Искра революции»
Союзполиграфпрома Государственного комитета СССР
по делам издательств, полиграфии и книжной торговли.
103001, Москва, Трехпрудный пер., 9.

С $\frac{1702070000-024}{056(01)-86}$ 146—86, ч. 2

Golovinà L.

ALGEBRA LINEAL Y SUS APLICACIONES

(3ª edición)

No obstante su pequeño volumen, el libro contiene los problemas fundamentales del curso de álgebra lineal, así como sus distintas aplicaciones, incluyendo la investigación de las curvas y superficies de segundo orden, la noción sobre tensores y otros problemas.

En el libro se exponen los conceptos primordiales referentes a los espacios lineales y euclidianos, y a las transformaciones lineales; además se estudian problemas sobre vectores y se obtiene la forma canónica de las matrices de las transformaciones autoconjugada y otrogonal en el espacio euclidiano, dándose ejemplos básicos de la teoría de las formas cuadráticas.

Como suplemento a las formas cuadráticas, se examina la teoría general de las líneas y superficies de segundo orden. Dos capítulos están consagrados a las transformaciones de Lorentz y nociones fundamentales de la teoría especial de la relatividad. En el capítulo sobre grupos, además de las definiciones principales, se incluye una selección de ejemplos.

Un mérito evidente del libro es la acertada elección del material, en el cual se han examinado problemas que no entran en el programa de los estudiantes de especialidades no matemáticas, pero que son de cierto interés para éstos. Con ello, las nociones indispensables previas y el nivel de la exposición son tales que, al leer el texto, los estudiantes no encuentran ninguna dificultad.

La obra está destinada a estudiantes y profesores de centros de enseñanza superior. También les es de gran utilidad a los ingenieros que deseen conocer las nociones fundamentales del álgebra lineal, empleando una fuente que no exige información previa de matemáticas superiores.

Kagán V.

LOBACHEVSKI

En este libro se narra de la vida y de las actividades sociales y científicas del eminente matemático ruso N. I. Lobachevski.

Se relata en detalle sobre la niñez y adolescencia del gran sabio, sobre su familia, profesores del liceo que influyeron en la formación del carácter y la concepción del mundo del científico. El autor presta especial atención a los años estudiantiles de Lobachevski en la Universidad de Kazán, y en particular, a sus estudios de las matemáticas.

Gran parte de la obra está dedicada a la actividad científica de Lobachevski. El lector conocerá el manual de estudio titulado "Geometría", donde por primera vez la geometría absoluta fue separada de la euclidiana. Se trata en detalle la creación de la geometría no euclidiana, se analiza el trabajo "Geometría, investigación de la teoría de las líneas paralelas". Un apartado importante del libro se destina a los años más fructíferos de la obra de Lobachevski (1827—1846), en que fuera rector de la Universidad de Kazán. En este período él publicó sus trabajos más notables: "Sobre los principios de la geometría", "Nuevos principios de la geometría con la teoría completa de las paralelas", etc., que son analizados. El siguiente apartado ha sido dedicado a la relación de sus contemporáneos hacia las ideas de Lobachevski, al desarrollo de estas ideas en los años ulteriores, a la aplicación de la geometría de Lobachevski a otras partes de las matemáticas, así como a la mecánica, física y cosmología.

Este libro resultará sin duda de interés a estudiantes, maestros, profesores y a toda persona aficionada a las matemáticas.